Necessary and Sufficient Conditions and a Provably Efficient Algorithm for Separable Topic Discovery

Weicong Ding,* Prakash Ishwar, *IEEE Senior Member* and Venkatesh Saligrama, *IEEE Senior Member*

***Abstract*—We develop necessary and sufficient conditions and a novel provably consistent and efficient algorithm for discovering topics (latent factors) from observations (documents) that are realized from a probabilistic mixture of shared latent factors that have certain properties. Our focus is on the class of topic models in which each shared latent factor contains a novel word that is unique to that factor, a property that has come to be known as separability. Our algorithm is based on the key insight that the novel words correspond to the extreme points of the convex hull formed by the row-vectors of a suitably normalized word co-occurrence matrix. We leverage this geometric insight to establish polynomial computation and sample complexity bounds based on a few isotropic random projections of the rows of the normalized word co-occurrence matrix. Our proposed random-projections-based algorithm is naturally amenable to an efficient distributed implementation and is attractive for modern web-scale distributed data mining applications.**

***Index Terms*—Topic Modeling, Separability, Random Projection, Solid Angle, Necessary and Sufficient Conditions.**

I. Introduction

Topic modeling refers to a family of generative models and associated algorithms for discovering the (latent) topical structure shared by a large corpus of documents. They are important for organizing, searching, and making sense of a large text corpus [1]. In this paper we describe a novel geometric approach, with provable statistical and computational efficiency guarantees, for learning the latent topics in a document collection. This work is a culmination of a series of recent publications on certain structure-leveraging methods for topic modeling with provable theoretical guarantees [2]–[5].

We consider a corpus of M documents, indexed by $m = 1,\ldots,M$, each composed of words from a fixed vocabulary of size W. The distinct words in the vocabulary are indexed by $w = 1,\ldots,W$. Each document m is viewed as an unordered "bag of words" and is represented by an empirical $W \times 1$ word-counts vector $\mathbf{X}^m$, where $X_{w,m}$ is the number of times that word w appears in document m [1], [5]–[7]. The entire document corpus is then represented by the $W \times M$ matrix $\mathbf{X} = \left[\mathbf{X}^1,\ldots,\mathbf{X}^M\right]$. [1] A "topic" is a $W \times 1$ distribution over the vocabulary. A topic model posits the existence of $K < \min(W,M)$ latent topics that are *shared* among all M documents in the corpus. The topics can be collectively represented by the K columns $\boldsymbol{\beta}^1,\ldots,\boldsymbol{\beta}^K$ of a $W \times K$ column-stochastic "topic matrix" $\boldsymbol{\beta}$. Each document m is conceptually modeled as being generated independently of all other documents through a two-step process: 1) first draw a $K \times 1$ document-specific distribution over topics $\boldsymbol{\theta}^m$ from a prior distribution $\Pr(\boldsymbol{\alpha})$ on the probability simplex with some hyper-parameters $\boldsymbol{\alpha}$; 2) then draw N iid words according to a $W \times 1$ document-specific word distribution over the vocabulary given by $\mathbf{A}^m = \sum_{k=1}^K \boldsymbol{\beta}^k \theta_{k,m}$ which is a convex combination (probabilistic mixture) of the latent topics. Our goal is to estimate $\boldsymbol{\beta}$ from the matrix of empirical observations $\mathbf{X}$. To appreciate the difficulty of the problem, consider a typical benchmark dataset such as a news article collection from the New York Times (NYT) [8] that we use in our experiments. In this dataset, after suitable pre-processing, $W = 14,943$, $M = 300,000$, and, on average, $N = 298$. Thus, $N \ll W \ll M$, $\mathbf{X}$ is very sparse, and M is very large. Typically, $K \approx 100 \ll \min(W,M)$.

This estimation problem in topic modeling has been extensively studied. The prevailing approach is to compute the MAP/ML estimate [1]. The true posterior of $\boldsymbol{\beta}$ given $\mathbf{X}$, however, is intractable to compute and the associated MAP and ML estimation problems are in fact NP-hard in the general case [9], [10]. This necessitates the use of sub-optimal methods based on approximations and heuristics such as Variational-Bayes and MCMC [6], [11]–[13]. While they produce impressive empirical results on many real-world datasets, guarantees of asymptotic consistency or efficiency for these approaches are either weak or non-existent. This makes it difficult to evaluate *model fidelity*: failure to produce satisfactory results in new datasets could be due to the use of approximations and heuristics or due to model mis-specification which is more fundamental. Furthermore, these sub-optimal approaches are computationally intensive for large text corpora [5], [7].

To overcome the hardness of the topic estimation problem in its full generality, a new approach has emerged to learn the topic model by imposing additional structure on the model parameters [3], [5], [7], [9], [14], [15]. This paper focuses on a key structural property of the topic matrix $\boldsymbol{\beta}$ called **topic separability** [3], [5], [7], [15] wherein every latent topic contains at least one word that is **novel** to it, i.e., the word is unique to that topic and is absent from the other topics. This is, in essence, a property of the support of the latent topic matrix $\boldsymbol{\beta}$. The topic separability property can be motivated by the fact that for many real-world datasets, the empirical topic estimates produced by popular Variational-Bayes and Gibbs Sampling approaches are approximately separable [5], [7]. Moreover, it has recently been shown that the separability property will be approximately satisfied with high probability when the dimension of the vocabulary W scales sufficiently faster than the number of topics K and $\boldsymbol{\beta}$ is a realization

[1]When it is clear from the context, we will use $X_{w,m}$ to represent either the empirical word-count or, by suitable column-normalization of $\mathbf{X}$, the empirical word-frequency.

of a Dirichlet prior that is typically used in practice [16]. Therefore, separability is a *natural* approximation for *most* high-dimensional topic models.

Our approach exploits the following geometric implication of the key separability structure. If we associate each word in the vocabulary with a row-vector of a suitably normalized empirical word co-occurrence matrix, **the set of novel words correspond to the extreme points** of the convex hull formed by the row-vectors of all words. We leverage this geometric insight and develop a provably consistent and efficient algorithm. Informally speaking, we establish the following result:

Theorem 1. *If the topic matrix is separable and the mixing weights satisfy a minimum information-theoretically necessary technical condition, then our proposed algorithm runs in* **polynomial time** *in* M, W, N, K*, and estimates the topic matrix* **consistently** *as* $M \to \infty$ *with* $N \geq 2$ *held fixed. Moreover, our proposed algorithm can estimate* $\boldsymbol{\beta}$ *to within an* ϵ *element-wise error with a probability at least* $1 - \delta$ *if* $M \geq$ ***Poly*** $(W, 1/N, K, \log(1/\delta), 1/\epsilon)$.

The asymptotic setting $M \to \infty$ with N held fixed is motivated by text corpora in which the number of words in a single document is small while the number of documents is large. We note that our algorithm can be applied to any family of topic models whose topic mixing weights prior $\Pr(\boldsymbol{\alpha})$ satisfies a minimum information-theoretically necessary technical condition. In contrast, the standard Bayesian approaches such as Variational-Bayes or MCMC need to be hand-designed separately for each specific topic mixing weights prior.

The highlight of our approach is to identify the novel words as extreme points through appropriately defined **random projections**. Specifically, we project the row-vector of each word in an appropriately normalized word co-occurrence matrix along a few independent and isotropically distributed random directions. The fraction of times that a word attains the maximum value along a random direction is a measure of its degree of robustness as an extreme point. This process of random projections followed by counting the number of times a word is a maximizer can be efficiently computed and is robust to the perturbations induced by sampling noise associated with having only a very small number of words per document N. In addition to being computationally efficient, it turns out that this random projections based approach (1) requires the *minimum* information-theoretically necessary technical conditions on the topic prior for asymptotic consistency, and (2) can be naturally parallelized and distributed. As a consequence, it can provably achieve the efficiency guarantees of a centralized method while requiring insignificant communication between *distributed* document collections [5]. This is attractive for web-scale topic modeling of large distributed text corpora.

Another advance of this paper is the identification of necessary and sufficient conditions on the mixing weights for consistent separable topic estimation. In previous work we showed that a *simplicial* condition on the mixing weights is both necessary and sufficient for consistently *detecting* all the novel words [4]. In this paper we complete the characterization by showing that an *affine independence* condition on the mixing weights is necessary and sufficient for consistently *estimating* a separable topic matrix. These conditions are satisfied by practical choices of topic priors such as the Dirichlet distribution [6]. All these necessary conditions are information-theoretic and algorithm-independent, i.e., they are irrespective of the specific statistics of the observations or the algorithms that are used. The provable statistical and computational efficiency guarantees of our proposed algorithm hold true under these necessary and sufficient conditions.

The rest of this paper is organized as follows. We review related work on topic modeling as well as the separability property in various domains in Sec. II. We introduce the separability property on $\boldsymbol{\beta}$, the simplicial and affine independence conditions on mixing weights, and the extreme point geometry that motivates our approach in Sec. III. We then discuss how the solid angle can be used to identify robust extreme points to deal with a finite number of samples (words per document) in Sec. IV. We describe our overall algorithm and sketch its analysis in Sec. V. We demonstrate the performance of our approach in Sec. VI on various synthetic and real-world examples. Proofs of all results appear in the appendices.

II. Related Work

The idea of modeling text documents as mixtures of a few semantic topics was first proposed in [17] where the mixing weights were assumed to be deterministic. Latent Dirichlet Allocation (LDA) in the seminal work of [6] extended this to a probabilistic setting by modeling topic mixing weights using Dirichlet priors. This setting has been further extended to include other topic priors such as the log-normal prior in the Correlated Topic Model [18]. LDA models and their derivatives have been successful on a wide range of problems in terms of achieving good empirical performance [1], [13].

The prevailing approaches for estimation and inference problems in topic modeling are based on MAP or ML estimation [1]. However, the computation of posterior distributions conditioned on observations $\mathbf{X}$ is intractable [6]. Moreover, the MAP estimation objective is non-convex and has been shown to be $\mathcal{NP}$-hard [9], [10]. Therefore various approximation and heuristic strategies have been employed. These approaches fall into two major categories – sampling approaches and optimization approaches. Most sampling approaches are based on Markov Chain Monte Carlo (MCMC) algorithms that seek to generate (approximately) independent samples from a Markov Chain that is carefully designed to ensure that the sample distribution converges to the true posterior [11], [19]. Optimization approaches are typically based on the so-called Variational-Bayes methods. These methods optimize the parameters of a simpler parametric distribution so that it is close to the true posterior in terms of KL divergence [6], [12]. Expectation-Maximization-type algorithms are typically used in these methods. In practice, while both Variational-Bayes and MCMC algorithms have similar performance, Variational-Bayes is typically faster than MCMC [1], [20].

Nonnegative Matrix Factorization (NMF) is an alternative approach for topic estimation. NMF-based methods exploit the fact that both the topic matrix $\boldsymbol{\beta}$ and the mixing weights are nonnegative and attempt to decompose the empirical

observation matrix $\mathbf{X}$ into a product of a nonnegative topic matrix $\boldsymbol{\beta}$ and the matrix of mixing weights by minimizing a cost function of the form [20]–[23]

$$\sum_{m=1}^{M} d(\mathbf{X}^m, \boldsymbol{\beta}\boldsymbol{\theta}^m) + \lambda\psi(\boldsymbol{\beta}, \boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^M),$$

where $d(,)$ is some measure of closeness and ψ is a regularization term which enforces desirable properties, e.g., sparsity, on $\boldsymbol{\beta}$ and the mixing weights. The NMF problem, however, is also known to be non-convex and $\mathcal{NP}$-hard [24] in general. Sub-optimal strategies such as alternating minimization, greedy gradient descent, and heuristics are used in practice [22].

In contrast to the above approaches, a new approach has recently emerged which is based on imposing additional structure on the model parameters [3], [5], [7], [9], [14], [15]. These approaches show that the topic discovery problem lends itself to provably consistent and polynomial-time solutions by making assumptions about the *structure* of the topic matrix $\boldsymbol{\beta}$ and the distribution of the mixing weights. In this category of approaches are methods based on a tensor decomposition of the moments of $\mathbf{X}$ [14], [25]. The algorithm in [25] uses second order empirical moments and is shown to be asymptotically consistent when the topic matrix $\boldsymbol{\beta}$ has a special sparsity structure. The algorithm in [14] uses the third order tensor of observations. It is, however, strongly tied to the specific structure of the Dirichlet prior on the mixing weights and requires knowledge of the concentration parameters of the Dirichlet distribution [14]. Furthermore, in practice these approaches are computationally intensive and require some initial coarse dimensionality reduction, gradient descent speedups, and GPU acceleration to process large-scale text corpora like the NYT dataset [14].

Our work falls into the family of approaches that exploit the separability property of $\boldsymbol{\beta}$ and its geometric implications [3], [5], [7], [9], [15], [26], [27]. An asymptotically consistent polynomial-time topic estimation algorithm was first proposed in [9]. However, this method requires solving W linear programs, each with W variables and is computationally impractical. Subsequent work improved the computational efficiency [15], [23], but theoretical guarantees of asymptotic consistency (when N fixed, and the number of documents $M \to \infty$) are unclear. Algorithms in [7] and [3] are both practical and provably consistent. Each requires a stronger and slightly different technical condition on the topic mixing weights than [9]. Specifically, [7] imposes a full-rank condition on the second-order correlation matrix of the mixing weights and proposes a Gram-Schmidt procedure to identify the extreme points. Similarly, [3] imposes a diagonal-dominance condition on the same second-order correlation matrix and proposes a random projections based approach. These approaches are tied to the specific conditions imposed and they both fail to detect all the novel words and estimate topics when the imposed conditions (which are sufficient but not necessary for consistent novel word detection or topic estimation) fail to hold in some examples [5]. The random projections based algorithm proposed in [5] is both practical and provably consistent. Furthermore, it requires fewer constraints on the topic mixing weights.

We note that the separability property has been exploited in other recent work as well [26], [27]. In [27], a singular value decomposition based approach is proposed for topic estimation. In [26], it is shown that the standard Variational-Bayes approximation can be asymptotically consistent if $\boldsymbol{\beta}$ is separable. However, the additional constraints proposed essentially boil down to the requirement that each document contain predominantly only one topic. In addition to assuming the existence of such "pure" documents, [26] also requires a strict initialization. It is thus unclear how this can be achieved using only the observations $\mathbf{X}$.

The separability property has been re-discovered and exploited in the literature across a number of different fields and has found application in several problems. To the best of our knowledge, this concept was first introduced as the *Pure Pixel Index* assumption in the Hyperspectral Image unmixing problem [28]. This work assumes the existence of pixels in a hyper-spectral image containing predominantly one species. Separability has also been studied in the NMF literature in the context of ensuring the uniqueness of NMF [29]. Subsequent work has led to the development of NMF algorithms that exploit separability [23], [30]. The uniqueness and correctness results in this line of work has primarily focused on the noiseless case. We finally note that separability has also been recently exploited in the problem of learning multiple ranking preferences from pairwise comparisons for personal recommendation systems and information retrieval [31], [32] and has led to provably consistent and efficient estimation algorithms.

III. Topic Separability, Necessary and Sufficient Conditions, and the Geometric Intuitions

In this section, we unravel the key ideas that motivate our algorithmic approach by focusing on the ideal case where there is no "sampling-noise", i.e., each document is infinitely long ($N = \infty$). In the next section, we will turn to the finite N case. We recall that $\boldsymbol{\beta}$ and $\mathbf{X}$ denote the $W \times K$ topic matrix and the $W \times M$ empirical word counts/frequency matrix respectively. Also, M, W, and K denote, respectively, the number of documents, the vocabulary size, and the number of topics. For convenience, we group the document-specific mixing weights, the $\boldsymbol{\theta}^m$'s, into a $K \times M$ weight matrix $\boldsymbol{\theta} = \left[\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^M\right]$ and the document-specific distributions, the $\mathbf{A}^m$'s, into a $W \times M$ document distribution matrix $\mathbf{A} = \left[\mathbf{A}^1, \ldots, \mathbf{A}^M\right]$. The generative procedure that describes a topic model then implies that $\mathbf{A} = \boldsymbol{\beta}\boldsymbol{\theta}$. In the ideal case considered in this section ($N = \infty$), the empirical word *frequency* matrix $\mathbf{X} = \mathbf{A}$. **Notation:** A vector $\mathbf{a}$ without specification will denote a column-vector, $\mathbf{1}$ the all-ones column vector of suitable size, $\mathbf{X}^i$ the i-th column vector and $\mathbf{X}_j$ the j-th row vector of matrix $\mathbf{X}$, and $\bar{\mathbf{B}}$ a suitably row-normalized version (described later) of a nonnegative matrix $\mathbf{B}$. Also, $[n] := \{1, \ldots, n\}$.

A. Key Structural Property: Topic Separability

We first introduce separability as a key structural property of a topic matrix $\boldsymbol{\beta}$. Formally,

Definition 1. (Separability) *A topic matrix $\boldsymbol{\beta} \in \mathbb{R}^{W \times K}$ is separable if for each topic k, there is some word i such that $\boldsymbol{\beta}_{i,k} > 0$ and $\boldsymbol{\beta}_{i,l} = 0$, $\forall\ l \neq k$.*

Topic separability implies that each topic contains word(s) which appear only in that topic. We refer to these words as the **novel words** of the K topics. Figure 1 shows an example of a separable $\boldsymbol{\beta}$ with $K = 3$ topics. Words 1 and 2 are novel to topic 1, words 3 and 4 to topic 2, and word 5 to topic 3. Other words that appear in multiple topics are called non-novel words (e.g., word 6). Identifying the novel words for K distinct topics is the key step of our proposed approach.

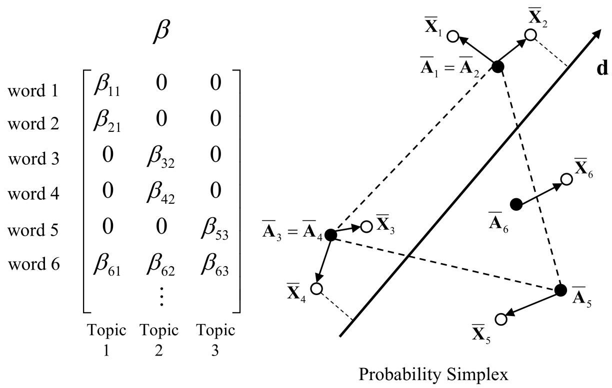


Fig. 1. An example of separable topic matrix $\boldsymbol{\beta}$ (left) and the underlying geometric structure (right) of the row space of the normalized document distribution matrix $\bar{\mathbf{A}}$. Note: the word ordering is only for visualization and has no bearing on separability. Solid circles represent **rows** of $\bar{\mathbf{A}}$. Empty circles represent **rows** of $\bar{\mathbf{X}}$ when N is finite (in the ideal case, $\bar{\mathbf{A}} = \bar{\mathbf{X}}$). Projections of $\bar{\mathbf{A}}_w$'s (resp. $\bar{\mathbf{X}}_w$'s) along a random isotropic direction $\mathbf{d}$ can be used to identify novel words.

We note that separability has been empirically observed to be approximately satisfied by topic estimates produced by Variational-Bayes and MCMC based algorithms [5], [7], [26]. More fundamentally, in very recent work [16], it has been shown that topic separability is an inevitable consequence of having a relatively small number of topics in a very large vocabulary (high-dimensionality). In particular, when the K columns (topics) of $\boldsymbol{\beta}$ are independently sampled from a Dirichlet distribution (on a $(W-1)$-dimensional probability simplex), the resulting topic matrix $\boldsymbol{\beta}$ will be (approximately) separable with probability tending to 1 as W scales to infinity sufficiently faster than K. A Dirichlet prior on $\boldsymbol{\beta}$ is widely-used in smoothed settings of topic modeling [1].

As we will discuss next in Sec. III-C, the topic separability property combined with additional conditions on the second-order statistics of the mixing weights leads to an intuitively appealing geometric property that can be exploited to develop a provably consistent and efficient topic estimation algorithm.

B. Conditions on the Topic Mixing Weights

Topic separability alone does not guarantee that there will be a unique $\boldsymbol{\beta}$ that is consistent with all the observations $\mathbf{X}$. This is illustrated in Fig. 2 [4]. Therefore, in an effort to develop provably consistent topic estimation algorithms, a number of different conditions have been imposed on the topic mixing weights $\boldsymbol{\theta}$ in the literature [3], [5], [7], [9], [15]. Complementing the work in [4] which identifies necessary and sufficient conditions for consistent *detection* of novel words, in this paper we identify necessary and sufficient conditions for consistent *estimation* of a separable topic matrix. Our necessity results are *information-theoretic and algorithm-independent* in nature, meaning that they are independent of any specific statistics of the observations and the algorithms used. The novel words and the topics can only be identified up to a permutation and this is accounted for in our results.

Let $\mathbf{a} := \mathbb{E}(\boldsymbol{\theta}^m)$ and $\mathbf{R} := \mathbb{E}(\boldsymbol{\theta}^m \boldsymbol{\theta}^{m\top})$ be the $K \times 1$ expectation vector and the $K \times K$ correlation matrix of the weight prior $\Pr(\boldsymbol{\alpha})$. Without loss of generality, we can assume that the elements of $\mathbf{a}$ are strictly positive since otherwise some topic(s) will not appear in the corpus. A key quantity is $\bar{\mathbf{R}} := \mathrm{diag}(\mathbf{a})^{-1}\mathbf{R}\,\mathrm{diag}(\mathbf{a})^{-1}$ which may be viewed as a "normalized" second-moment matrix of the weight vector. The following conditions are central to our results.

Condition 1. (Simplicial Condition) *A matrix $\mathbf{B}$ is (row-wise) γ_s-simplicial if any row-vector of $\mathbf{B}$ is at a Euclidean distance of at least $\gamma_s > 0$ from the convex hull of the remaining row-vectors. A topic model is γ_s-simplicial if its normalized second-moment $\bar{\mathbf{R}}$ is γ_s-simplicial.*

Condition 2. (Affine-Independence) *A matrix $\mathbf{B}$ is (row-wise) γ_a-affine-independent if $\min_{\boldsymbol{\lambda}} \| \sum_{k=1}^{K} \lambda_k \mathbf{B}_k \|_2 / \|\boldsymbol{\lambda}\|_2 \geq \gamma_a > 0$, where $\mathbf{B}_k$ is the k-th row of $\mathbf{B}$ and the minimum is over all $\boldsymbol{\lambda} \in \mathbb{R}^K$ such that $\boldsymbol{\lambda} \neq \mathbf{0}$ and $\sum_{k=1}^{K} \lambda_k = 0$. A topic model is γ_a-affine-independent if its normalized second-moment $\bar{\mathbf{R}}$ is γ_a-affine-independent.*

Here, γ_s and γ_a are called the simplicial and affine-independence constants respectively. They are condition numbers which measure the degree to which the conditions that they are respectively associated with hold. The larger that these condition numbers are, the easier it is to estimate the topic matrix. Going forward, we will say that a matrix is simplicial (resp. affine independent) if it is γ_s-simplicial (resp. γ_a-affine-independent) for some $\gamma_s > 0$ (resp. $\gamma_a > 0$). The simplicial condition was first proposed in [9] and then further investigated in [4]. This paper is the first to identify affine-independence as both *necessary and sufficient* for consistent separable topic estimation. Before we discuss their geometric implications, we point out that affine-independence is stronger than the simplicial condition:

Proposition 1. *$\bar{\mathbf{R}}$ is γ_a-affine-independent $\Rightarrow$ $\bar{\mathbf{R}}$ is at least γ_a-simplicial. The reverse implication is false in general.*

The Simplicial Condition is both Necessary and Sufficient for Novel Word Detection: We first focus on detecting all the novel words of the K distinct topics. For this task, the simplicial condition is an algorithm-independent, information-theoretic necessary condition. Formally,

Lemma 1. *(Simplicial Condition is Necessary for Novel Word Detection [4, Lemma 1]) Let $\boldsymbol{\beta}$ be separable and $W > K$. If there exists an algorithm that can consistently identify all novel words of all K topics from $\mathbf{X}$, then $\bar{\mathbf{R}}$ is simplicial.*

The key insight behind this result is that when $\bar{\mathbf{R}}$ is non-

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ & \ldots & \end{pmatrix}}_{\boldsymbol{\beta}^{(1)}} \underbrace{\begin{pmatrix} \leftarrow & \boldsymbol{\theta}_1 & \rightarrow \\ \leftarrow & \boldsymbol{\theta}_2 & \rightarrow \\ \leftarrow & 0.5\boldsymbol{\theta}_1 + 0.5\boldsymbol{\theta}_2 & \rightarrow \end{pmatrix}}_{\boldsymbol{\theta}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.5 & 0.5 & 0 \\ & \ldots & \end{pmatrix}}_{\boldsymbol{\beta}^{(2)}} \underbrace{\begin{pmatrix} \leftarrow & \boldsymbol{\theta}_1 & \rightarrow \\ \leftarrow & \boldsymbol{\theta}_2 & \rightarrow \\ \leftarrow & 0.5\boldsymbol{\theta}_1 + 0.5\boldsymbol{\theta}_2 & \rightarrow \end{pmatrix}}_{\boldsymbol{\theta}}$$

Fig. 2. Example showing that topic separability **alone** does not guarantee a unique solution to the problem of estimating $\boldsymbol{\beta}$ from $\mathbf{X}$. Here, $\boldsymbol{\beta}_1\boldsymbol{\theta} = \boldsymbol{\beta}_2\boldsymbol{\theta} = \mathbf{A}$ is a document distribution matrix that is consistent with two different topic matrices $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ that are both separable.

simplicial, we can construct two distinct separable topic matrices with different sets of novel words which induce the same distribution on the empirical observations $\mathbf{X}$. Geometrically, the simplicial condition guarantees that the K rows of $\bar{\mathbf{R}}$ will be extreme points of the convex hull that they themselves form. Therefore, if $\bar{\mathbf{R}}$ is not simplicial, there will exist at least one redundant topic which is just a convex combination of the other topics.

It turns out that $\bar{\mathbf{R}}$ being simplicial is also sufficient for consistent novel word detection. This is a direct consequence of the consistency guarantees of our approach as outlined in Theorem 3.

Affine-Independence is Necessary and Sufficient for Separable Topic Estimation: We now focus on estimating a separable topic matrix $\boldsymbol{\beta}$, which is a stronger requirement than detecting novel words. It naturally requires conditions that are stronger than the simplicial condition. Affine-independence turns out to be an algorithm-independent, information-theoretic necessary condition. Formally,

Lemma 2. *(Affine-Independence is Necessary for Separable Topic Estimation) Let $\boldsymbol{\beta}$ be separable with $W \geq 2+K$. If there exists an algorithm that can consistently estimate $\boldsymbol{\beta}$ from $\mathbf{X}$, then its normalized second-moment $\bar{\mathbf{R}}$ is affine-independent.*

Similar to Lemma 1, if $\bar{\mathbf{R}}$ is not affine-independent, we can construct two distinct and separable topic matrices that induce the same distribution on the observation which makes consistent topic estimation impossible. Geometrically, every point in a convex set can be decomposed *uniquely* as a convex combination of its extreme points, if, and only if, the extreme points are affine-independent. Hence, if $\bar{\mathbf{R}}$ is not affine-independent, a non-novel word can be assigned to different subsets of topics.

The sufficiency of the affine-independence condition in separable topic estimation is again a direct consequence of the consistency guarantees of our approach as in Theorems 3 and 4. We note that since affine-independence implies the simplicial condition (Proposition 1), affine-independence is sufficient for novel word detection as well.

Connection to Other Conditions on the Mixing Weights: We briefly discuss other conditions on the mixing weights $\boldsymbol{\theta}$ that have been exploited in the literature. In [7], [15], $\mathbf{R}$ (equivalently $\bar{\mathbf{R}}$) is assumed to have full-rank (with minimum eigenvalue $\gamma_r > 0$). In [3], $\bar{\mathbf{R}}$ is assumed to be diagonal-dominant, i.e., $\forall i, j, i \neq j, \bar{\mathbf{R}}_{i,i} - \bar{\mathbf{R}}_{i,j} \geq \gamma_d > 0$. They are both sufficient conditions for detecting all the novel words of all distinct topics. The constants γ_r and γ_d are condition numbers which measure the degree to which the full-rank and diagonal-dominance conditions hold respectively. They are counterparts of γ_s and γ_a and like them, the larger they are, the easier it is to consistently detect the novel words and estimate $\boldsymbol{\beta}$. The relationships between these conditions are summarized in Proposition 2 and illustrated in Fig. 3.

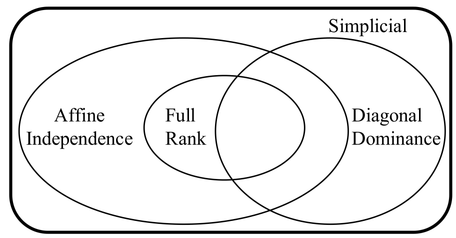


Fig. 3. Relationships between Simplicial, Affine-Independence, Full Rank, and Diagonal Dominance conditions on the normalized second-moment $\bar{\mathbf{R}}$.

Proposition 2. *Let $\bar{\mathbf{R}}$ be the normalized second-moment of the topic prior. Then,*

1) *$\bar{\mathbf{R}}$ is full rank with minimum eigenvalue $\gamma_r \Rightarrow \bar{\mathbf{R}}$ is at least γ_r-affine-independent $\Rightarrow \bar{\mathbf{R}}$ is at least γ_r-simplicial.*
2) *$\bar{\mathbf{R}}$ is γ_d-diagonal-dominant $\Rightarrow \bar{\mathbf{R}}$ is at least γ_d-simplicial.*
3) *$\bar{\mathbf{R}}$ being diagonal-dominant neither implies nor is implied by $\bar{\mathbf{R}}$ being affine-independent (or full-rank).*

We note that in our earlier work [5], the provable guarantees for estimating the separable topic matrix require $\bar{\mathbf{R}}$ to have full rank. The analysis in this paper provably extends the guarantees to the affine-independence condition.

C. Geometric Implications and Random Projections Based Algorithm

We now demonstrate the geometric implications of topic separability combined with the simplicial/ affine-independence condition on the topic mixing weights. To highlight the key ideas we focus on the ideal case where $N = \infty$. Then, the empirical document word-frequency matrix $\mathbf{X} = \mathbf{A} = \boldsymbol{\beta}\boldsymbol{\theta}$.

Novel Words are Extreme Points: To expose the underlying geometry, we normalize the rows of $\mathbf{A}$ and $\boldsymbol{\theta}$ to obtain row-stochastic matrices $\bar{\mathbf{A}} := \mathrm{diag}(\mathbf{A}\mathbf{1})^{-1}\mathbf{A}$ and $\bar{\boldsymbol{\theta}} := \mathrm{diag}(\boldsymbol{\theta}\mathbf{1})^{-1}\boldsymbol{\theta}$. Then since $\mathbf{A} = \boldsymbol{\beta}\boldsymbol{\theta}$, we have $\bar{\mathbf{A}} = \bar{\beta}\bar{\theta}$ where $\bar{\boldsymbol{\beta}} := \mathrm{diag}(\mathbf{A}\mathbf{1})^{-1}\boldsymbol{\beta}\,\mathrm{diag}(\boldsymbol{\theta}\mathbf{1})$ is a row-normalized "topic matrix" which is both row-stochastic and separable with the same sets of novel words as $\boldsymbol{\beta}$.

Now consider the row vectors of $\bar{\mathbf{A}}$ and $\bar{\boldsymbol{\theta}}$. First, it can be shown that if $\bar{\mathbf{R}}$ is simplicial (cf. Condition 1) then, with high probability, no row of $\bar{\boldsymbol{\theta}}$ will be in the convex hull of

the others (see Appendix D). Next, the separability property ensures that if w is a novel word of topic k, then $\bar{\beta}_{wk} = 1$ and $\bar{\beta}_{wj} = 0 \; \forall j \neq k$ so that $\bar{\mathbf{A}}_w = \bar{\boldsymbol{\theta}}_k$. Revisiting the example in Fig. 1, the rows of $\bar{\mathbf{A}}$ which correspond to novel words, e.g., words 1 through 5, are all row-vectors of $\bar{\boldsymbol{\theta}}$ and together form a convex hull of K extreme points. For example, $\bar{\mathbf{A}}_1 = \bar{\mathbf{A}}_2 = \bar{\boldsymbol{\theta}}_1$ and $\bar{\mathbf{A}}_3 = \bar{\mathbf{A}}_4 = \bar{\boldsymbol{\theta}}_2$. If, however, w is a non-novel word, then $\bar{\mathbf{A}}_w = \sum_k \bar{\beta}_{wk} \bar{\boldsymbol{\theta}}_k$ lives inside the convex hull of the rows of $\bar{\boldsymbol{\theta}}$. In Fig. 1, row $\bar{\mathbf{A}}_6$ which corresponds to non-novel word 6, is inside the convex hull of $\bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2, \bar{\boldsymbol{\theta}}_3$. In summary, the novel words can be detected as extreme points of all the row-vectors of $\bar{\mathbf{A}}$. Also, multiple novel words of the same topic correspond to the same extreme point (e.g., $\bar{\mathbf{A}}_1 = \bar{\mathbf{A}}_2 = \bar{\boldsymbol{\theta}}_1$). Formally,

Lemma 3. *Let $\bar{\mathbf{R}}$ be γ_s simplicial and $\boldsymbol{\beta}$ be separable. Then, with probability at least $1-2K\exp(-c_1 M)-\exp(-c_2 M)$, the i-th row of $\bar{\mathbf{A}}$ is an extreme point of the convex hull spanned by all the rows of $\bar{\mathbf{A}}$ if, and only if, word i is novel. Here the constant $c_1 := \gamma_s^2 a_{\min}^4 / 4\lambda_{\max}$ and $c_2 := \gamma_s^4 a_{\min}^4 / 2\lambda_{\max}^2$. The model parameters are defined as follows. $a_{\min}$ is the minimum element of $\mathbf{a}$. $\lambda_{\max}$ is the maximum singular-value of $\bar{\mathbf{R}}$.*

To see how identifying novel words can help us estimate $\boldsymbol{\beta}$, recall that the row-vectors of $\bar{\mathbf{A}}$ corresponding to novel words coincide with the rows of $\bar{\boldsymbol{\theta}}$. Thus $\bar{\boldsymbol{\theta}}$ is known once one novel word for each topic is known. Also, for all words w, $\bar{\mathbf{A}}_w = \sum_k \bar{\beta}_{wk} \bar{\boldsymbol{\theta}}_k$. Thus, if we can *uniquely* decompose $\bar{\mathbf{A}}_w$ as a convex combination of the extreme points, then the coefficients of the decomposition will give us the w-th row of $\bar{\boldsymbol{\beta}}$. A unique decomposition exists with high probability when $\bar{\mathbf{R}}$ is affine-independent and can be found by solving a constrained linear regression problem. This gives us $\bar{\boldsymbol{\beta}}$. Finally, noting that $\operatorname{diag}(\mathbf{A}\mathbf{1})\bar{\boldsymbol{\beta}} = \boldsymbol{\beta}\operatorname{diag}(\boldsymbol{\theta}\mathbf{1})$, $\boldsymbol{\beta}$ can be recovered by suitably renormalizing rows and then columns of $\bar{\boldsymbol{\beta}}$. To sum up,

Lemma 4. *Let $\mathbf{A}$ and one novel word per distinct topic be given. If $\bar{\mathbf{R}}$ is γ_a affine-independent, then, with probability at least $1-2K\exp(-c_1 M)-\exp(-c_2 M)$, $\boldsymbol{\beta}$ can be recovered uniquely via constrained linear regression. Here the constant $c_1 := \gamma_a^2 a_{\min}^4 / 4\lambda_{\max}$ and $c_2 := \gamma_a^4 a_{\min}^4 / 2\lambda_{\max}^2$. The model parameters are defined as follows. $a_{\min}$ is the minimum element of $\mathbf{a}$. $\lambda_{\max}$ is the maximum singular-value of $\bar{\mathbf{R}}$.*

Lemmas 3 and 4 together provide a geometric approach for learning $\boldsymbol{\beta}$ from $\mathbf{A}$ (equivalently $\bar{\mathbf{A}}$): (1) Find extreme points of rows of $\bar{\mathbf{A}}$. Cluster the rows of $\bar{\mathbf{A}}$ that correspond to the same extreme point into the same group. (2) Express the remaining rows of $\bar{\mathbf{A}}$ as convex combinations of the K distinct extreme points. (3) Renormalize $\bar{\boldsymbol{\beta}}$ to obtain $\boldsymbol{\beta}$.

Detecting Extreme Points using Random Projections: A key contribution of our approach is an efficient random projections based algorithm to detect novel words as extreme points. The idea is illustrated in Fig. 1: if we project every point of a convex body onto an isotropically distributed random direction $\mathbf{d}$, the maximum (or minimum) projection value must correspond to one of the extreme points with probability 1. On the other hand, the non-novel words will not have the maximum projection value along any random direction. Therefore, by repeatedly projecting all the points onto a few isotropically distributed random directions, we can detect all the extreme points with very high probability as the number of random directions increase. An explicit bound on the number of projections needed appears in Theorem 3.

Finite N in Practice: The geometric intuition discussed above was based on the row-vectors of $\bar{\mathbf{A}}$. When $N = \infty$, $\bar{\mathbf{A}} = \bar{\mathbf{X}}$ the matrix of row-normalized empirical word-frequencies of all documents. If N is finite but very large, $\bar{\mathbf{A}}$ can be well-approximated by $\bar{\mathbf{X}}$ thanks to the law of large numbers. However, in real-word text corpora, $N \ll W$ (e.g., $N = 298$ while $W = 14,943$ in the NYT dataset). Therefore, the row-vectors of $\bar{\mathbf{X}}$ are significantly perturbed away from the ideal rows of $\bar{\mathbf{A}}$ as illustrated in Fig. 1. We discuss the effect of small N and how we address the accompanying issues next.

IV. Topic Geometry with Finite Samples: Word Co-occurrence Matrix Representation, Solid Angle, and Random Projections based approach

The extreme point geometry sketched in Sec. III-C is perturbed when N is small as highlighted in Fig. 1. Specifically, the rows of the empirical word-frequency matrix $\mathbf{X}$ deviate from the rows of $\mathbf{A}$. This creates several problems: (1) points in the convex hull corresponding to non-novel words may also become "outlier" extreme points (e.g., $\bar{\mathbf{X}}_6$ in Fig. 1); (2) some extreme points that correspond to novel words may no longer be extreme (e.g., $\bar{\mathbf{X}}_3$ in Fig. 1); (3) multiple novel words corresponding to the same extreme point may become multiple distinct extreme points (e.g., $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ in Fig. 1). Unfortunately, these issues do not vanish as M increases with N fixed – a regime which captures the characteristics of typical benchmark datasets – because the dimensionality of the rows (equal to M) also increases. There is no "averaging" effect to smoothen-out the sampling noise.

Our solution is to seek a new representation, a statistic of $\mathbf{X}$, which can not only smoothen out the sampling noise of individual documents, but also preserve the same extreme point geometry induced by the separability and affine independence conditions. In addition, we also develop an extreme point robustness measure that naturally arises within our random projections based framework. This robustness measure can be used to detect and exclude the "outlier" extreme points.

A. Normalized Word Co-occurrence Matrix Representation

We construct a suitably normalized word co-occurrence matrix from $\mathbf{X}$ as our new representation. The co-occurrence matrix converges almost surely to an ideal statistic as $M \to \infty$ for any fixed $N \geq 2$. Simultaneously, in the asymptotic limit, the original novel words continue to correspond to extreme points in the new representation and overall extreme point geometry is preserved.

The new representation is (conceptually) constructed as follows. First randomly divide all the words in each document into two equal-sized independent halves and obtain two $W \times K$ empirical word-frequency matrices $\mathbf{X}$ and $\mathbf{X}'$ each containing $N/2$ words. Then normalize their rows like in Sec. III-C to

obtain $\bar{\mathbf{X}}$ and $\bar{\mathbf{X}}'$ which are row-stochastic. The empirical word co-occurrence matrix of size $W \times W$ is then given by

$$\widehat{\mathbf{E}} := M\bar{\mathbf{X}}'\bar{\mathbf{X}}^\top \tag{1}$$

We note that in our random projection based approach, $\widehat{\mathbf{E}}$ is not *explicitly* constructed by multiplying $\bar{\mathbf{X}}'$ and $\bar{\mathbf{X}}$. Instead, we keep $\bar{\mathbf{X}}'$ and $\bar{\mathbf{X}}$ and exploit their sparsity properties to reduce the computational complexity of all subsequent processing.

Asymptotic Consistency: The first nice property of the word co-occurrence representation is its asymptotic consistency when N is fixed. As the number of documents $M \to \infty$, the empirical $\widehat{\mathbf{E}}$ converges, almost surely, to an ideal word co-occurrence matrix $\mathbf{E}$ of size $W \times W$. Formally,

Lemma 5. *([32, Lemma 2]) Let $\widehat{\mathbf{E}}$ be the empirical word co-occurrence matrix defined in Eq.* (1)*. Then,*

$$\widehat{\mathbf{E}} \xrightarrow[\textit{almost surely}]{M\to\infty} \bar{\boldsymbol{\beta}}\bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top =: \mathbf{E} \tag{2}$$

where $\bar{\boldsymbol{\beta}} := \operatorname{diag}^{-1}(\boldsymbol{\beta}\mathbf{a})\boldsymbol{\beta}\operatorname{diag}(\mathbf{a})$ *and* $\bar{\mathbf{R}} := \operatorname{diag}^{-1}(\mathbf{a})\mathbf{R}\operatorname{diag}^{-1}(\mathbf{a})$. *Furthermore, if* $\eta := \min_{1\le i\le W}(\boldsymbol{\beta}\mathbf{a})_i > 0$*, then* $\Pr(\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \ge \epsilon) \le 8W^2\exp(-\epsilon^2\eta^4 MN/20)$.

Here $\bar{\mathbf{R}}$ is the same normalized second-moment of the topic priors as defined in Sec. III and $\bar{\boldsymbol{\beta}}$ is a row-normalized version of $\boldsymbol{\beta}$. We make note of the abuse of notion for $\bar{\boldsymbol{\beta}}$ which was defined in Sec. III-C. It can be shown that the $\bar{\boldsymbol{\beta}}$ defined in Lemma 5 is the limit of the one defined in Sec. III-C as $M \to \infty$. The convergence result in Lemma 5 shows that the word co-occurrence representation $\mathbf{E}$ can be consistently estimated by $\widehat{\mathbf{E}}$ as $M \to \infty$ and the deviation vanishes exponentially in M which is large in typical benchmark datasets.

Novel Words are Extreme Points: Another reason for using this word co-occurrence representation is that it preserves the extreme point geometry. Consider the ideal word co-occurrence matrix $\mathbf{E} = \bar{\boldsymbol{\beta}}(\bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top)$. It is straightforward to show that if $\bar{\boldsymbol{\beta}}$ is separable and $\bar{\mathbf{R}}$ is simplicial then $(\bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top)$ is also simplicial. Using these facts it is possible to establish the following counterpart of Lemma 3 for $\mathbf{E}$:

Lemma 6. *(Novel Words are Extreme Points [5, Lemma 1]) Let $\bar{\mathbf{R}}$ be simplicial and $\boldsymbol{\beta}$ be separable. Then, a word i is novel if, and only if, the i-th row of $\mathbf{E}$ is an extreme point of the convex hull spanned by all the rows of $\mathbf{E}$.*

In another words, the novel words correspond to the extreme points of all the row-vectors of the ideal word co-occurrence matrix $\mathbf{E}$. Consider the example in Fig. 4 which is based on the same topic matrix $\boldsymbol{\beta}$ as in Fig. 1. Here, $\mathbf{E}_1 = \mathbf{E}_2, \mathbf{E}_3 = \mathbf{E}_4$, and $\mathbf{E}_5$ are $K = 3$ distinct extreme points of all row-vectors of $\mathbf{E}$ and $\mathbf{E}_6$, which corresponds to a non-novel word, is inside the convex hull.

Once the novel words are detected as extreme points, we can follow the same procedure as in Lemma 4 and express each row $\mathbf{E}_w$ of $\mathbf{E}$ as a unique convex combination of the K extreme rows of $\mathbf{E}$ or equivalently the rows of $(\bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top)$. The weights of the convex combination are the $\bar{\beta}_{wk}$'s. We can then apply the same row and column renormalization to obtain $\boldsymbol{\beta}$. The following result is the counterpart of Lemma 4 for $\mathbf{E}$:

Lemma 7. *Let $\mathbf{E}$ and one novel word for each distinct topic be given. If $\bar{\mathbf{R}}$ is affine-independent, then $\boldsymbol{\beta}$ can be recovered uniquely via constrained linear regression.*

One can follow the same steps as in the proof of Lemma 4. The only additional step is to check that $\bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top = [\bar{\mathbf{R}}, \bar{\mathbf{R}}\mathbf{B}]$ is affine-independent if $\bar{\mathbf{R}}$ is affine-independent.

We note that the finite sampling noise perturbation $\widehat{\mathbf{E}} - \mathbf{E}$ is still not 0 but vanishes as $M \to \infty$ (in contrast to the $\bar{\mathbf{X}}$ representation in Sec. III-C). However, there is still a possibility of observing "outlier" extreme points if a non-novel word lies on the facet of the convex hull of the rows of $\mathbf{E}$. We next introduce an extreme point robustness measure based on a certain *solid angle* that naturally arises in our random projections based approach, and discuss how it can be used to detect and distinguish between "true" novel words and such "outlier" extreme points.

B. Solid Angle Extreme Point Robustness Measure

To handle the impact of a small but nonzero perturbation $\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty$, we develop an extreme point "robustness" measure. This is necessary for not only applying our approach to real-world data but also to establish finite sample complexity bounds. Intuitively, a robustness measure should be able to distinguish between the "true" extreme points (row vectors that are novel words) and the "outlier" extreme points (row vectors of non-novel words that become extreme points due to the nonzero perturbation). Towards this goal, we leverage a key geometric quantity, namely, the *Normalized Solid Angle* subtended by the convex hull of the rows of $\mathbf{E}$ at an extreme point. To visualize this quantity, we revisit our running example in Fig. 4 and indicate the solid angles attached to each extreme point by the shaded regions. It turns out that this geometric quantity naturally arises in the context of random projections that was discussed earlier. To see this connection, in Fig. 4 observe that the shaded region attached to any extreme point

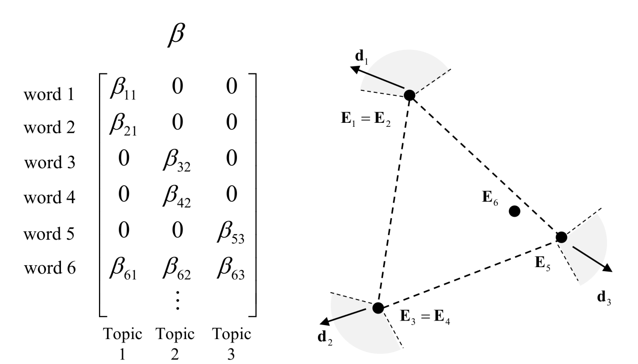


Fig. 4. An example of separable topic matrix $\boldsymbol{\beta}$ (left) and the underlying geometric structure (right) in the word co-occurrence representation. Note: the word ordering is only for visualization and has no bearing on separability. The example topic matrix $\boldsymbol{\beta}$ is the same as in Fig. 1. Solid circles represent the **rows** of $\mathbf{E}$. The shaded regions depict the solid angles subtended by each extreme point. $\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3$ are isotropic random directions along which each extreme point has maximum projection value. They can be used to estimate the solid angles.

coincides precisely with the set of directions along which its projection is larger (taking sign into account) than that of any other point (whether extreme or not). For example, in Fig. 4 the projection of $\mathbf{E}_1 = \mathbf{E}_2$ along $\mathbf{d}_1$ is larger than that of any other point. Thus, the solid angle attached to a point $\mathbf{E}_i$ (whether extreme or not) can be formally defined as the set of directions $\{\mathbf{d} : \forall j : \mathbf{E}_j \neq \mathbf{E}_1, \langle \mathbf{E}_i, \mathbf{d} \rangle > \langle \mathbf{E}_j, \mathbf{d} \rangle\}$. This set is nonempty only for extreme points. The solid angle defined above is a set. To derive a scalar robustness measure from this set and tie it to the idea of random projections, we adopt a statistical perspective and define the normalized solid angle of a point as the *probability* that the point will have the maximum projection value along an isotropically distributed random direction. Concretely, for the i-th word (row vector), the normalized solid angle q_i is defined as

$$q_i := \Pr(\forall j : \mathbf{E}_j \neq \mathbf{E}_i, \langle \mathbf{E}_i, \mathbf{d} \rangle > \langle \mathbf{E}_j, \mathbf{d} \rangle) \tag{3}$$

where $\mathbf{d}$ is drawn from an isotropic distribution in $\mathbb{R}^W$ such as the spherical Gaussian. The condition $\mathbf{E}_i \neq \mathbf{E}_j$ in Eq. (3) is introduced to exclude the multiple novel words of the same topic that correspond to the same extreme point. For instance, in Fig. 4 $\mathbf{E}_1 = \mathbf{E}_2$, Hence, for q_1, $j = 2$ is excluded. To make it practical to handle finite sample estimation noise we replace the condition $\mathbf{E}_j \neq \mathbf{E}_i$ by the condition $\|\mathbf{E}_i - \mathbf{E}_j\| \geq \zeta$ for some suitably defined ζ.

As illustrated in Fig. 4, the solid angle for all the extreme points are strictly positive given $\bar{\mathbf{R}}$ is γ_s-simplicial. On the other hand, for i that is non-novel, the corresponding solid angle q_i is zero by definition. Hence the extreme point geometry in Lemma 6 can be re-expressed in term of solid angles as follows:

Lemma 8. *(Novel Words have Positive Solid Angles) Let $\bar{\mathbf{R}}$ be simplicial and $\boldsymbol{\beta}$ be separable. Then, word i is a novel word if, and only if, $q_i > 0$.*

We denote the smallest solid angle among the K distinct extreme points by $q_\wedge > 0$. This is a robust condition number of the convex hull formed by the rows of $\mathbf{E}$ and is related to the simplicial constant γ_s of $\bar{\mathbf{R}}$.

In a real-world dataset we have access to only an empirical estimate $\widehat{\mathbf{E}}$ of the ideal word co-occurrence matrix $\mathbf{E}$. If we replace $\mathbf{E}$ with $\widehat{\mathbf{E}}$, then the resulting empirical solid angle estimate $\widehat{q}_i$ will be very close to the ideal q_i if $\widehat{\mathbf{E}}$ is close enough to $\mathbf{E}$. Then, the solid angles of "outlier" extreme points will be close to 0 while they will be bounded away from zero for the "true" extreme points. One can then hope to correctly identify all K extreme points by *rank-ordering* all empirical solid angle estimates and selecting the K distinct row-vectors that have the largest solid angles. This forms the basis of our proposed algorithm. The problem now boils down to efficiently estimating the solid angles and establishing the asymptotic convergence of the estimates as $M \to \infty$. We next discuss how random projections can be used to achieve these goals.

C. Efficient Solid Angle Estimation via Random Projections

The definition of the normalized solid angle in Eq. (3) motivates an efficient algorithm based on *random projections* to estimate it. For convenience, we first rewrite Eq. (3) as

$$q_i = \mathbb{E}\left[\mathbb{I}\{\forall j : \|\mathbf{E}_j - \mathbf{E}_i\| \geq \zeta,\ \mathbf{E}_i\mathbf{d} \geq \mathbf{E}_j\mathbf{d}\}\right] \tag{4}$$

and then propose to estimate it by

$$\hat{q}_i = \frac{1}{P}\sum_{r=1}^{P} \mathbb{I}(\forall j : \widehat{E}_{i,i} + \widehat{E}_{j,j} - 2\widehat{E}_{i,j} \geq \zeta/2, \quad \widehat{\mathbf{E}}_i\mathbf{d}^r > \widehat{\mathbf{E}}_j\mathbf{d}^r) \tag{5}$$

where $\mathbf{d}^1, \ldots, \mathbf{d}^P \in \mathbf{R}^{W\times 1}$ are P iid directions drawn from an isotropic distribution in $\mathbf{R}^W$. Algorithmically, by Eq. (5), we approximate the solid angle q_i at the i-th word (row-vector) by first projecting all the row-vectors onto P iid isotropic random directions and then calculating the fraction of times each row-vector achieves the maximum projection value. It turns out that the condition $\widehat{E}_{i,i} + \widehat{E}_{j,j} - 2\widehat{E}_{i,j} \geq \zeta/2$ is equivalent to $\|\mathbf{E}_i - \mathbf{E}_j\| \geq \zeta$ in terms of its ability to exclude multiple novel words from the same topic and is adopted for its simplicity. [2]

This procedure of taking random projections followed by calculating the number of times a word is a maximizer via Eq. (5) provides a consistent estimate of the solid angle in Eq. (3) as $M \to \infty$ and the number of projections P increases. The high-level idea is simple: as P increases, the empirical average in Eq. 5 converges to the corresponding expectation. Simultaneously, as M increases, $\widehat{\mathbf{E}} \xrightarrow{a.s.} \mathbf{E}$. Overall, the approximation $\widehat{q}_i$ proposed in Eq (5) using random projections converges to q_i.

This random projections based approach is also computationally efficient for the following reasons. First, it enables us to avoid the explicit construction of the $W \times W$ dimensional matrix $\widehat{\mathbf{E}}$: Recall that each column of $\mathbf{X}$ and $\mathbf{X}'$ has no more than $N \ll W$ nonzero entries. Hence $\mathbf{X}$ and $\mathbf{X}'$ are both sparse. Since $\widehat{\mathbf{E}}\mathbf{d} = M\bar{\mathbf{X}}'(\bar{\mathbf{X}}^\top\mathbf{d})$, the projection can be calculated using two sparse matrix-vector multiplications. Second, it turns out that the number of projections P needed to guarantee consistency is small. In fact in Theorem 3 we provide a sufficient upper bound for P which is a polynomial function of $\log(W)$, $\log(1/\delta)$ and other model parameters, where δ is the probability that the algorithm fails to detect all the distinct novel words.

Parallelization, Distributed and Online Settings: Another advantage of the proposed random projections based approach is that it can be *parallelized* and is naturally amenable to *online* or *distributed* settings. This is based on the following observation that each projection has an additive structure:

$$\widehat{\mathbf{E}}\mathbf{d}^r = M\bar{\mathbf{X}}'\bar{\mathbf{X}}^\top\mathbf{d}^r = M\sum_{m=1}^{M}\bar{\mathbf{X}}^{m\prime}\bar{\mathbf{X}}^{m\top}\mathbf{d}^r.$$

The P projections can also be computed independently. Therefore,

- In a *distributed* setting in which the documents are stored on distributed servers, we can first share the same random

[2] We abuse the symbol ζ by using it to indicate different thresholds in these conditions.

directions across servers and then aggregate the projection values. The communication cost is only the "partial" projection values and is therefore insignificant [5] and does not scale as the number of observations N, M increases.

- In an *online* setting in which the documents are streamed in an online fashion [20], we only need to keep all the projection values and update the projection values (hence the empirical solid angle estimates) when new documents arrive.

The additive and independent structure guarantees that the statistical efficiency of these variations are the same as the centralized "batch" implementation. For the rest of this paper, we only focus on the centralized version.

Outline of Overall Approach: Our overall approach can be summarized as follows. (1) Estimate the empirical solid angles using P iid isotropic random directions as in Eq. 5. (2) Select the K words with distinct word co-occurrence patterns (rows) that have the largest empirical solid angles. (3) Estimate the topic matrix using constrained linear regression as in Lemma 4. We will discuss the details of our overall approach in the next section and establish guarantees for its computational and statistical efficiency.

V. Algorithm and Analysis

Algorithm 1 describes the main steps of our overall random projectons based algorithm which we call RP. The two main steps, novel word detection and topic matrix estimation are outlined in Algorithms 2 and 3 respectively. Algorithm 2 outlines the random projection and rank-ordering steps. Algorithm 3 describes the constrained linear regression and the renormalization steps in a combined way.

Algorithm 1 RP

Input: Text documents $\bar{\mathbf{X}}$, $\bar{\mathbf{X}}'(W \times M)$; Number of topics K; Number of iid random projections P; Tolerance parameters $\zeta, \epsilon > 0$.
Output: Estimate of the topic matrix $\widehat{\boldsymbol{\beta}}(W \times K)$.
Set of Novel Words $\mathcal{I} \leftarrow$NovelWordDetect($\bar{\mathbf{X}}, \bar{\mathbf{X}}', K, P, \zeta$)
$\hat{\boldsymbol{\beta}} \leftarrow$EstimateTopics($\mathcal{I}, \bar{\mathbf{X}}, \bar{\mathbf{X}}', \epsilon$)

Computational Efficiency: We first summarize the computational efficiency of Algorithm 1:

Theorem 2. *Let the number of novel words for each topic be a constant relative to M, W, N. Then, the running time of Algorithm 1 is $\mathcal{O}(MNP + WP + WK^3)$.*

This efficiency is achieved by exploiting the sparsity of $\mathbf{X}$ and the property that there are only a small number of novel words in a typical vocabulary. A detailed analysis of the computational complexity is presented in the appendix. Here we point out that in order to upper bound the computation time of the linear regression in Algorithm 3 we used $\mathbf{O}(WK^3)$ for W matrix inversions, one for each of the words in the vocabulary. In practice, a gradient descent implementation can be used for the constrained linear regression which is much more efficient. We also note that these W optimization problems are decoupled given the set of detected novel words. Therefore, they can be parallelized in a straightforward manner [5].

Algorithm 2 NovelWordDetect (via Random Projections)

Input: $\bar{\mathbf{X}}, \bar{\mathbf{X}}'$; Number of topics K; Number of projections P; Tolerance ζ;
Output: The set of all novel words of K distinct topics $\mathcal{I}$.
$\hat{q}_i \leftarrow 0,\ \forall i = 1, \dots, W, \quad \widehat{\mathbf{E}} \leftarrow M\bar{\mathbf{X}}'\bar{\mathbf{X}}^\top$.
for all $r = 1, \dots, P$ **do**
  Sample $\mathbf{d}^r \in \mathbb{R}^W$ from an isotropic prior.
  $\mathbf{v} \leftarrow M\bar{\mathbf{X}}'\bar{\mathbf{X}}^\top \mathbf{d}^r$
  $i^* \leftarrow \arg\max_{1 \le i \le W} \mathbf{v}_i, \quad \hat{q}_{i^*} \leftarrow \hat{q}_{i^*} + 1/P$
  $\hat{J}_{i^*} \leftarrow \{j : \widehat{E}_{i^*,i^*} + \widehat{E}_{j,j} - 2\widehat{E}_{i^*,j} \ge \zeta/2\}$
  for all $k \in \hat{J}^c_{i^*}$ **do**
    $\hat{J}_k \leftarrow \{j : \widehat{E}_{k,k} + \widehat{E}_{j,j} - 2\widehat{E}_{k,j} \ge \zeta/2\}$
    if $\{\forall j \in \hat{J}_k, v_k > v_j\}$ **then**
      $\hat{q}_k \leftarrow \hat{q}_k + 1/P$
    end if
  end for
end for
$\mathcal{I} \leftarrow \emptyset$, $k \leftarrow 0$, $j \leftarrow 1$
while $k < K$ **do**
  $i \leftarrow$ index of the j^{th} largest value of $\{\hat{q}_1, \dots, \hat{q}_W\}$.
  if $\{\forall p \in \mathcal{I}, \widehat{E}_{p,p} + \widehat{E}_{i,i} - 2\widehat{E}_{i,p} \ge \zeta/2\}$ **then**
    $\mathcal{I} \leftarrow \mathcal{I} \cup \{i\}$, $k \leftarrow k + 1$
  end if
  $j \leftarrow j + 1$
end while
Return $\mathcal{I}$.

Algorithm 3 EstimateTopics

Input: $\mathcal{I} = \{i_1, \dots, i_K\}$ set of novel words, one for each of the K topics; $\widehat{\mathbf{E}}$; precision parameter ϵ
Output: $\widehat{\boldsymbol{\beta}}$, which is the estimate of the $\boldsymbol{\beta}$ matrix
$\widehat{\mathbf{E}}^*_w = \left[\widehat{\mathbf{E}}_{w,i_1}, \dots, \widehat{\mathbf{E}}_{w,i_K}\right]$
$\mathbf{Y} = (\widehat{\mathbf{E}}^{*\top}_{i_1}, \dots, \widehat{\mathbf{E}}^{*\top}_{i_K})^\top$
for all $i = 1, \dots, W$ **do**
  Solve $\mathbf{b}^* := \arg\min_{\mathbf{b}} \|\widehat{\mathbf{E}}^*_i - \mathbf{b}\mathbf{Y}\|^2$
  subject to $b_j \ge 0, \sum_{j=1}^K b_j = 1$
  using precision ϵ for the stopping-criterion.
  $\widehat{\boldsymbol{\beta}}_i \leftarrow (\frac{1}{M}\mathbf{X}_i \mathbf{1})\mathbf{b}^*$
end for
$\widehat{\boldsymbol{\beta}} \leftarrow$column normalize $\widehat{\boldsymbol{\beta}}$

Asymptotic Consistency and Statistical Efficiency: We now summarize the asymptotic consistency and sample complexity bounds for Algorithm 1. The analysis is a combination of the consistency of the novel word detection step (Algorithm 2) and the topic estimation step (Algorithm 3). We state the results for both of these steps. First, for detecting all the novel words of the K distinct topics, we have the following result:

Theorem 3. *Let topic matrix $\boldsymbol{\beta}$ be separable and $\bar{\mathbf{R}}$ be γ-simplicial. If the projection directions are iid sampled from any isotropic distribution, then Algorithm 2 can identify all*

the novel words of the K distinct topics as $M, P \to \infty$. Furthermore, $\forall \delta \geq 0$, if

$$M \geq 20\frac{\log(2W/\delta)}{N\rho^2\eta^4} \text{ and } P \geq 8\frac{\log(2W/\delta)}{q_\wedge^2} \tag{6}$$

then Algorithm 2 fails with probability at most δ. The model parameters are defined as follows. $\rho = \min\{\frac{d}{8}, \frac{\pi d_2 q_\wedge}{4W^{1.5}}\}$ where $d = (1-b)^2\gamma^2/\lambda_{\max}$, $d_2 \triangleq (1-b)\gamma$, $\lambda_{\max}$ is the maximum eigenvalue of $\bar{\mathbf{R}}$, $b = \max_{j \in \mathcal{C}_0, k} \bar{\beta}_{j,k}$, and $\mathcal{C}_0$ is the set of non-novel words. Finally, $q_\wedge$ is the minimum solid angle of the extreme points of the convex hull of the rows of $\mathbf{E}$.

The detailed proof is presented in the appendix. The results in Eq. (6) provide a sufficient finite sample complexity bound for novel word detection. The bound is *polynomial* with respect to M, W, K, N, $\log(\delta)$ and other model parameters. The number of projections P that impacts the computational complexity scales as $\log(W)/q_\wedge^2$ in this sufficient bound where $q_\wedge$ can be upper bounded by $1/K$. In practice, we have found that setting $P = \mathcal{O}(K)$ is a good choice [5].

We note that the result in Theorem 3 only requires the simplicial condition which is the *minimum* condition required for consistent novel word detection (Lemma 1). This theorem holds true if the topic prior $\bar{\mathbf{R}}$ satisfies stronger conditions such as affine-independence. We also point out that our proof in this paper holds for *any isotropic distribution* on the random projection directions $\mathbf{d}^1, \ldots, \mathbf{d}^P$. The previous result in [5], however, only applies to some specific isotopic distributions such as the Spherical Gaussian or the uniform distribution in a unit ball. In practice, we use Spherical Gaussian since sampling from such prior is simple and requires only $\mathcal{O}(W)$ time for generating each random direction.

Next, given the successful detection of the set of novel words for all topics, we have the following result for the accurate estimation of the separable topic matrix $\boldsymbol{\beta}$:

Theorem 4. *Let topic matrix $\boldsymbol{\beta}$ be separable and $\bar{\mathbf{R}}$ be γ_a-affine-independent. Given the successful detection of novel words for all K distinct topics, the output of Algorithm 3 $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ element-wise (up to a column permutation). Specifically, if*

$$M \geq \frac{2560W^2K\log(W^4K/\delta)}{N\gamma_a^2 a_{\min}^2\eta^4\epsilon^2} \tag{7}$$

then $\forall i, k$, $\widehat{\beta}_{i,k}$ will be ϵ close to $\beta_{i,k}$ with probability at least $1-\delta$, for any $0 < \epsilon < 1$. η is the same as in Theorem 3. $a_{\min}$ is the minimum value in $\mathbf{a}$.

We note that the sufficient sample complexity bound in Eq. (7) is again polynomial in terms of all the model parameters. Here we only require $\bar{\mathbf{R}}$ to be affine-independent. Combining Theorem 3 and Theorem 4 gives the consistency and sample complexity bounds of our overall approach in Algorithm 1.

VI. Experimental Results

In this section, we present experimental results on both synthetic and real world datasets. We report different performance measures that have been commonly used in the topic modeling literature. When the ground truth is available (Sec. VI-A), we use the ℓ_1 *reconstruction error* between the ground truth topics and the estimates after proper topic alignment. For the real-world text corpus in Sec. VI-B, we report the *held-out probability*, which is a standard measure used in the topic modeling literature. We also *qualitatively* (semantically) compare the topics extracted by the different approaches using the top probable words for each topic.

A. Semi-synthetic text corpus

In order to validate our proposed algorithm, we generate "semi-synthetic" text corpora by sampling from a synthetic, yet realistic, ground truth topic model. To ensure that the semi-synthetic data is similar to real-world data, in terms of dimensionality, sparsity, and other characteristics, we use the following generative procedure adapted from [5], [7].

We first train an LDA model (with $K = 100$) on a real-world dataset using a standard Gibbs Sampling method with default parameters (as described in [11], [33]) to obtain a topic matrix $\boldsymbol{\beta}_0$ of size $W \times K$. The real-world dataset that we use to generate our synthetic data is derived from a New York Times (NYT) articles dataset [8]. The original vocabulary is first pruned based on document frequencies. Specifically, as is standard practice, only words that appear in more than 500 documents are retained. Thereafter, again as per standard practice, the words in the so-called stop-word list are deleted as recommended in [34]. After these steps, $M = 300,000$, $W = 14,943$, and the average document length $N = 298$. We then generate semi-synthetic datasets, for various values of M, by fixing $N = 300$ and using $\boldsymbol{\beta}_0$ and a Dirichlet topic prior. As suggested in [11] and used in [5], [7], we use symmetric hyper-parameters (0.03) for the Dirichlet topic prior.

The $W \times K$ topic matrix $\boldsymbol{\beta}_0$ may not be separable. To enforce separability, we create a new *separable* $(W+K) \times K$ dimensional topic matrix $\boldsymbol{\beta}_{\text{sep}}$ by inserting K synthetic novel words (one per topic) having suitable probabilities in each topic. Specifically, $\boldsymbol{\beta}_{\text{sep}}$ is constructed by transforming $\boldsymbol{\beta}_0$ as follows. First, for each synthetic novel word in $\boldsymbol{\beta}_{\text{sep}}$, the value of the sole nonzero entry in its row is set to the probability of the most probable word in the topic (column) of $\boldsymbol{\beta}_0$ for which it is a novel word. Then the resulting $(W+K) \times K$ dimensional nonnegative matrix is renormalized column-wise to make it column-stochastic. Finally, we generate semi-synthetic datasets, for various values of M, by fixing $N = 300$ and using $\boldsymbol{\beta}_{\text{sep}}$ and the same symmetric Dirichlet topic prior used for $\boldsymbol{\beta}_0$.

We use the name *Semi-Syn* to refer to datasets that are generated using $\boldsymbol{\beta}_0$ and the name *Semi-Syn+Novel* for datasets generated using $\boldsymbol{\beta}_{\text{sep}}$.

In our proposed random projections based algorithm, which we call RP, we set $P = 150 \times K$, $\zeta = 0.05$, and $\epsilon = 10^{-4}$. We compare RP against the provably efficient algorithm RecoverL2 in [7] and the standard Gibbs Sampling based LDA algorithm (denoted by Gibbs) in [11], [33]. In order to measure the performance of different algorithms in our experiments based on semi-synthetic data, we compute the ℓ_1 norm of the *reconstruction error* between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. Since

all column permutations of a given topic matrix correspond to the same topic model (for a corresponding permutation of the topic mixing weights), we use a bipartite graph matching algorithm to optimally match the columns of $\widehat{\boldsymbol{\beta}}$ with those of $\boldsymbol{\beta}$ (based on minimizing the sum of ℓ_1 distances between all pairs of matching columns) before computing the ℓ_1 norm of the reconstruction error between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$.

The results on both *Semi-Syn+Novel* NYT and *Semi-Syn* NYT are summarized in Fig. 5 for all three algorithms for various choices of the number of documents M. We note that in these figures the ℓ_1 norm of the error has been normalized by the number of topics ($K = 100$).

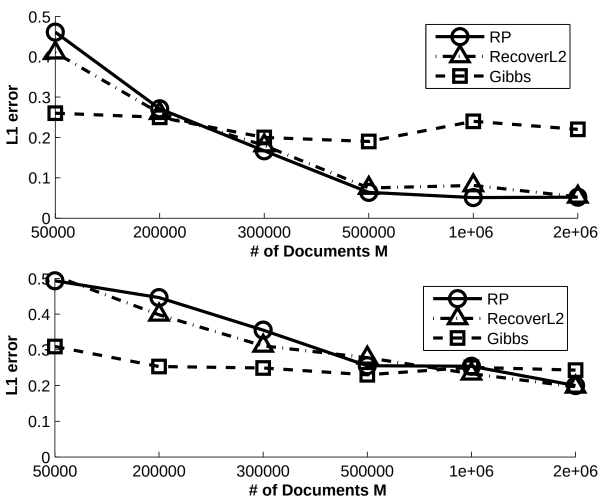


Fig. 5. ℓ_1 norm of the error in estimating the topic matrix $\boldsymbol{\beta}$ for various M ($K = 100$): (Top) *Semi-Syn+Novel* NYT; (Bottom) *Semi-Syn* NYT. RP is the proposed algorithm, RecoverL2 is a provably efficient algorithm from [7], and Gibbs is the Gibbs Sampling approximation algorithm in [11]. In RP, $P = 150K$, $\zeta = 0.05$, and $\epsilon = 10^-4$.

As Fig. 5 shows, when the separability condition is strictly satisfied (*Semi-Syn+Novel*), the reconstruction error of RP converges to 0 as M becomes large and outperforms the approximation-based Gibbs. When the separability condition is not strictly satisfied (*Semi-Syn*), the reconstruction error of RP is comparable to Gibbs (a practical benchmark).

Solid Angle and Model Selection: In our proposed algorithm RP, the number of topics K (the model-order) needs to be specified. When K is unavailable, it needs to be estimated from the data. Although not the focus of this work, Algorithm 2, which identifies novel words by sorting and clustering the estimated solid angles of words, can be suitably modified to estimate K.

Indeed, in the ideal scenario where there is no sampling noise ($M = \infty, \widehat{\mathbf{E}} = \mathbf{E}$, and $\forall i, \hat{q}_i = q_i$), only novel words have positive solid angles ($\hat{q}_i$'s) and the rows of $\widehat{\mathbf{E}}$ corresponding to the novel words of the same topic are identical, i.e., the distance between the rows is zero or, equivalently, they are within a neighborhood of size zero of each other. Thus, the number of distinct neighborhoods of size zero among the nonzero solid angle words equals K.

In the nonideal case M is finite. If M is sufficiently large, one can expect that the estimated solid angles of non-novel words will not all be zero. They are, however, likely to be much smaller than those of novel words. Thus to reliably estimate K one should not only exclude words with exactly zero solid angle estimates, but also those above some nonzero threshold. When M is finite, the the rows of $\widehat{\mathbf{E}}$ corresponding to the novel words of the same topic are unlikely to be identical, but if M is sufficiently large they are likely to be close to each other. Thus, if the threshold ζ in Algorithm 2, which determines the size of the neighborhood for clustering all novel words belonging to the same topic, is made sufficiently small, then each neighborhood will have only novel words belonging to the same topic.

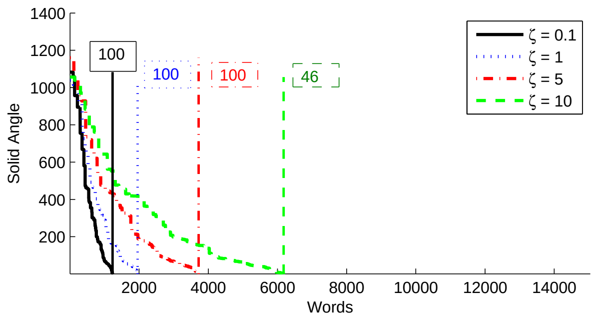
With the two modifications discussed above, the number of distinct neighborhoods of a suitably nonzero size (determined by $\zeta > 0$) among the words whose solid angle estimates are larger than some threshold $\tau > 0$ will provide an estimate of K. The values of τ and ζ should, in principle, decrease to zero as M increases to infinity. Leaving the task of unraveling the dependence of τ and ζ on M to future work, here we only provide a brief empirical validation on both the *Semi-Syn+Novel* and *Semi-Syn* NYT datasets. We set $M = 2,000,000$ so that the reconstruction error has essentially converged (see Fig. 5), and consider different choices of the threshold ζ.

We run Algorithm 2 with $K = 100$, $P = 150 \times K$, and a new line of code: 16': (**if** $\{\hat{q}_i = 0\}$, **break**); inserted between lines 16 and 17 (this corresponds to $\tau = 0$). The input hyperparameter $K = 100$ is not the actual number of estimated topics. It should be interpreted as specifying an upper bound on the number of topics. The value of (little) k when Algorithm 2 terminates (see lines 14–21) provides an estimate of the number of topics.

Figure 6 illustrates how the solid angles of all words, sorted in descending order, decay for different choices of ζ and how they can be used to detect the novel words and estimate the value of K. We note that in both the semi-synthetic datasets, for a wide range of values of ζ (0.1–5), the modified Algorithm 2 correctly estimates the value of K as 100. When ζ is large (e.g., $\zeta = 10$ in Fig. 6), many interior points would be declared as novel words and multiple ideal novel words would be grouped into one cluster resulting. This causes K to be underestimated (46 and 41 in Fig. 6).

B. Real-world data

We now describe results on the actual real-world NYT dataset that was used in Sec. VI-A to construct the semi-synthetic datasets. Since ground truth topics are unavailable, we measure performance using the so-called *predictive held-out log-probability*. This is a standard measure which is typically used to evaluate how well a learned topic model fits real-world data. To calculate this for each of the three topic estimation methods (Gibbs [11], [33], RecoverL2 [7], and RP), we first randomly select $60,000$ documents to test the goodness of fit and use the remaining $240,000$ documents to produce an estimate $\widehat{\boldsymbol{\beta}}$ of the topic matrix. Next we assume a Dirichlet prior on the topics and estimate its concentration hyper-parameter $\boldsymbol{\alpha}$. In Gibbs, this estimate $\widehat{\boldsymbol{\alpha}}$ is a byproduct of the algorithm. In RecoverL2 and RP this can be estimated



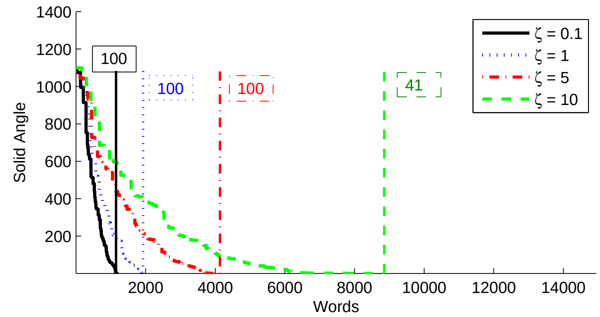


Fig. 6. Solid-angles (in descending order) of all $14943 + 100$ words in the *Semi-Syn+Sep* NYT dataset (left) and all 14943 words in the *Semi-Syn* NYT dataset (right) estimated (for different values of ζ) by Algorithm 2 with $K = 100$, $P = 150 \times K$, $M = 2,000,000$, and a new line of code: 16': (**if** $\{\hat{q}_i = 0\}$, **break**); inserted between lines 16 and 17. The values of j and (little) k when Algorithm 2 terminates are indicated, respectively, by the position of the vertical dashed line and the rectangular box next to it for different ζ.

from $\widehat{\boldsymbol{\beta}}$ and $\mathbf{X}$. We then calculate the probability of observing the test documents given the learned topic model $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$:

$$\log \Pr(\mathbf{X}_{\text{test}} | \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}})$$

Since an exact evaluation of this predictive log-likelihood is intractable in general, we calculate it using the MCMC based approximation proposed in [19] which is now a standard approximation tool [33]. For RP, we use $P = 150 \times K$, $\zeta = 0.05$, and $\epsilon = 10^-4$ as in Sec. VI-A. We report the held-out log probability, normalized by the total number of words in the test documents, averaged across 5 training/testing splits. The results are summarized in Table I. As shown in Table I,

TABLE I
NORMALIZED HELD-OUT LOG PROBABILITY OF RP, RECOVERL2, AND GIBBS SAMPLING ON NYT TEST DATA. THE MEAN±STD'S ARE CALCULATED FROM 5 DIFFERENT RANDOM TRAINING-TESTING SPLITS.

K	RecoverL2	Gibbs	RP
50	-8.22±0.56	-7.42±0.45	-8.54±0.52
100	-7.63±0.52	-7.50±0.47	-7.45±0.51
150	-8.03±0.38	-7.31±0.41	-7.84±0.48
200	-7.85±0.40	-7.34±0.44	-7.69±0.42

Gibbs has the best descriptive power for new documents. RP and RecoverL2 have similar, but somewhat lower values than Gibbs. This may be attributed to missing novel words that appear only in the test set and are crucial to the success of RecoverL2 and RP. Specifically, in real-world examples, there is a model-mismatch as a result of which the data likelihoods of RP and RecoverL2 suffer.

Finally, we *qualitatively* access the topics produced by our RP algorithm. We show some example topics extracted by RP trained on the *entire* NYT dataset of $M = 300,000$ documents in Table II [3] For each topic, its most frequent words are listed. As can be seen, the estimated topics do form recognizable themes that can be assigned meaningful labels. The full list of all $K = 100$ topics estimated on the NYT dataset can be found in [3].

[3]The zzz prefix in the NYT vocabulary is used to annotate certain special named entities. For example, zzz_nfl annotates NFL.

TABLE II
EXAMPLES OF TOPICS ESTIMATED BY RP ON NYT

Topic label	Words in decreasing order of estimated probabilities
"weather"	weather wind air storm rain cold
"feeling"	feeling sense love character heart emotion
"election"	election zzz_florida ballot vote zzz_al_gore recount
"game"	yard game team season play zzz_nfl

VII. CONCLUSION AND DISCUSSION

This paper proposed a provably consistent and efficient algorithm for topic discovery. We considered a natural structural property – topic separability – on the topic matrix and exploited its geometric implications. We resolved the necessary and sufficient conditions that can guarantee consistent novel words detection as well as separable topic estimation. We then proposed a random projections based algorithm that has not only provably polynomial statistical and computational complexity but also state-of-the-art performance on semi-synthetic and real-world datasets.

While we focused on the standard centralized batch implementation in this paper, it turns out that our random projections based scheme is naturally amenable to an efficient distributed implementation which is of interest when the documents are stored on a network of distributed servers. This is because the iid isotropic projection directions can be precomputed and shared across document servers, and counts, projections, and co-occurrence matrix computations have an additive structure which allows partial computations to be performed at each document server locally and then aggregated at a fusion center with only a small communication cost. It turns out that the distributed implementation can provably match the polynomial computational and statistical efficiency guarantees of its centralized counterpart. As a consequence, it provides a provably efficient alternative to the distributed topic estimation problem which has been tackled using variations of MCMC or Variational-Bayes in the literature [20], [35]–[37] This is appealing for modern web-scale databases, e.g., those generated by Twitter Streaming. A comprehensive theoretical and empirical investigation of the distributed variation of our algorithm can be found in [5].

Separability of general measures: We defined and studied the notion of separability for a $W \times K$ topic matrix $\boldsymbol{\beta}$ which is a finite collection of K probability distributions over a finite set (of size W). It turns out that we can extend the notion separability to a finite collection of measures over a measurable space. This necessitates making a small technical modification to the definition of separability to accommodate the possibility of only having "novel subsets" that have zero measure. We also show that our generalized definition of separability is equivalent to the so-called **irreducibility** property of a finite collection of measures that has recently been studied in the context of mixture models to establish conditions for the identifiability of the mixing components [38], [39].

Consider a collection of K measures $\nu_1, \ldots, \nu_K$ over a measurable space $(\mathcal{X}, \mathcal{F})$, where $\mathcal{X}$ is a set and $\mathcal{F}$ is a σ-algebra over $\mathcal{X}$. We define the generalized notion of separability for measures as follows.

Definition 2. (Separability) *A collection of K measures $\nu_1, \ldots, \nu_K$ over a measurable space $(\mathcal{X}, \mathcal{F})$ is separable if for all $k = 1, \ldots, K$,*

$$\inf_{A \in \mathcal{F}:\ \nu_k > 0} \max_{j:\ j \neq k} \frac{\nu_j(A)}{\nu_k(A)} = 0. \tag{8}$$

Separability requires that for each measure ν_k, there exists a sequence of measurable sets $A_n^{(k)}$, of nonzero measure with respect to ν_k, such that, for all $j \neq k$, the ratios $\nu_j(A_n^{(k)})/\nu_k(A_n^{(k)})$ vanish asymptotically. Intuitively, this means that for each measure there exists a sequence of nonzero-measure measurable subsets that are asymptotically "novel" for that measure. When $\mathcal{X}$ is a finite set as in topic modeling, this reduces to the existence of novel words as in Definition 1 and $A_n^{(k)}$ are simply the sets of novel words for topic k.

The separability property just defined is equivalent to the so-called irreducibility property. Informally, a collection of measures is irreducible if *only nonnegative linear combinations of them can produce a measure*. Formally,

Definition 3. (Irreducibility) *A collection of K measures $\nu_1, \ldots, \nu_K$ over a measurable space $(\mathcal{X}, \mathcal{F})$ is irreducible if the following condition holds: If $\forall A \in \mathcal{F}$, $\sum_{k=1}^{K} c_k \nu_k(A) \geq 0$, then for all $k = 1, \ldots, K$, $c_k \geq 0$.*

For a collection of nonzero measures,[4] these two properties are equivalent. Formally,

Lemma 9. *A collection of nonzero measures $\nu_1, \ldots, \nu_K$ over a measurable space $(\mathcal{X}, \mathcal{F})$ is irreducible if and only if it is separable. In particular, a topic matrix $\boldsymbol{\beta}$ is irreducible if and only if it is separable.*

The proof appears in Appendix M.

Topic models like LDA discussed in this paper belong to the larger family of Mixed Membership Latent Variable Models [13] which have been successfully employed in a variety of problems that include text analysis, genetic analysis, network community detection, and ranking and preference discovery. The structure-leveraging approach proposed in this paper can be potentially extended to this larger family of models. Some initial steps in this direction for rank and preference data are explored in [32].

Finally, in this entire paper, the topic matrix is assumed to be separable. While *exact* separability may be an idealization, as shown in [16], approximate separability is both theoretically inevitable and practically encountered when $W \gg K$. Extending the results of this work to approximately separable topic matrices is an interesting direction for future work. Some steps in this direction are explored in [40] in the context of learning mixed membership Mallows models for rankings.

[4] A measure ν is nonzero if there exists at least one measurable set A for which $\nu(A) > 0$.

Acknowledgment

This article is based upon work supported by the U.S. AFOSR under award number # FA9550-10-1-0458 (subaward # A1795) and the U.S. NSF under award numbers # 1527618 and # 1218992. The views and conclusions contained in this article are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the agencies.

Appendix

A. Proof of Lemma 1

Proof. The proof is by contradiction. We will show that if $\bar{\mathbf{R}}$ is non-simplicial, we can construct two topic matrices $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ whose sets of novel words are not identical and yet $\mathbf{X}$ has the same distribution under both models. The difference between constructed $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ is not a result of column permutation. This will imply the impossibility of consistent novel word detection.

Suppose $\bar{\mathbf{R}}$ is non-simplicial. Then we can assume, without loss of generality, that its first row is within the convex hull of the remaining rows, i.e., $\bar{\mathbf{R}}_1 = \sum_{j=2}^{K} c_j \bar{\mathbf{R}}_j$, where $\bar{\mathbf{R}}_j$ denotes the j-th row of $\bar{\mathbf{R}}$, and $c_2, \ldots, c_K \geq 0$, $\sum_{j=2}^{K} c_j = 1$ are convex combination weights. Compactly, $\mathbf{e}^\top \bar{\mathbf{R}} \mathbf{e} = 0$ where $\mathbf{e} := [-1, c_2, \ldots, c_K]^\top$. Recalling that $\bar{\mathbf{R}} = \mathrm{diag}(\mathbf{a})^{-1} \mathbf{R}\, \mathrm{diag}(\mathbf{a})^{-1}$, where $\mathbf{a}$ is a positive vector and $\mathbf{R} = \mathbb{E}(\boldsymbol{\theta}^m \boldsymbol{\theta}^{m\top})$ by definition, we have

$$\begin{aligned} 0 &= \mathbf{e}^\top \bar{\mathbf{R}} \mathbf{e} = (\mathrm{diag}(\mathbf{a})^{-1}\mathbf{e})^\top \mathbb{E}(\boldsymbol{\theta}^m \boldsymbol{\theta}^{m\top})(\mathrm{diag}(\mathbf{a})^{-1}\mathbf{e}) \\ &= \mathbb{E}(\|\boldsymbol{\theta}^{m\top} \mathrm{diag}(\mathbf{a})^{-1}\mathbf{e}\|_2^2), \end{aligned}$$

which implies that $\boldsymbol{\theta}^{m\top} \mathrm{diag}(\mathbf{a})^{-1}\mathbf{e} \overset{a.s.}{=} 0$. From this it follows that if we define two nonnegative row vectors $\mathbf{b}_1 := b\left[a_1^{-1}, 0, \ldots, 0\right]$ and $\mathbf{b}_2 = b\left[(1-\alpha)a_1^{-1}, \alpha c_2 a_2^{-1}, \ldots, \alpha c_K a_K^{-1}\right]$, where $b > 0, 0 < \alpha < 1$ are constants, then $\mathbf{b}_1 \boldsymbol{\theta}^m \overset{a.s.}{=} \mathbf{b}_2 \boldsymbol{\theta}^m$ for any distribution on $\boldsymbol{\theta}^m$.

Now we construct two separable topic matrices $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ as follows. Let $\mathbf{b}_1$ be the first row and $\mathbf{b}_2$ be the second in $\boldsymbol{\beta}^{(1)}$. Let $\mathbf{b}_2$ be the first row and $\mathbf{b}_1$ the second in $\boldsymbol{\beta}^{(2)}$. Let $\mathbf{B} \in \mathbb{R}^{W-2 \times K}$ be a valid separable topic matrix. Set the remaining $(W-2)$ rows of both $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ to be $\mathbf{B}(I_K - \mathrm{diag}(\mathbf{b}_1 + \mathbf{b}_2))$. We can choose b to be small enough to ensure that each element of $(\mathbf{b}_1 + \mathbf{b}_2)$ is strictly less than 1. This will ensure that $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ are column-stochastic

and therefore valid separable topic matrices. Observe that $\mathbf{b}_2$ has at lease two nonzero components. Thus, word 1 is novel for $\boldsymbol{\beta}^{(1)}$ but non-novel for $\boldsymbol{\beta}^{(2)}$.

By construction, $\boldsymbol{\beta}^{(1)}\boldsymbol{\theta} \overset{a.s.}{=} \boldsymbol{\beta}^{(2)}\boldsymbol{\theta}$, i.e., the distribution of $\mathbf{X}$ conditioned on $\boldsymbol{\theta}$ is the same for both models. Marginalizing over $\boldsymbol{\theta}$, the distribution of $\mathbf{X}$ under each topic matrix is the same. Thus no algorithm can consistently distinguish between $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ based on $\mathbf{X}$. □

B. Proof of Lemma 2

Proof. The proof is by contradiction. Suppose that $\bar{\mathbf{R}}$ is not affine-independent. Then there exists a $\boldsymbol{\lambda} \neq \mathbf{0}$ with $\mathbf{1}^\top\boldsymbol{\lambda} = 0$ such that $\boldsymbol{\lambda}^\top\bar{\mathbf{R}} = \mathbf{0}$ so that $\boldsymbol{\lambda}^\top\bar{\mathbf{R}}\boldsymbol{\lambda} = 0$. Recalling that $\bar{\mathbf{R}} = \operatorname{diag}(\mathbf{a})^{-1}\mathbf{R}\operatorname{diag}(\mathbf{a})^{-1}$, we have,

$$\begin{aligned} 0 &= \boldsymbol{\lambda}^\top\bar{\mathbf{R}}\boldsymbol{\lambda} = (\operatorname{diag}(\mathbf{a})^{-1}\boldsymbol{\lambda})^\top\,\mathbb{E}(\boldsymbol{\theta}^m\boldsymbol{\theta}^{m\top})(\operatorname{diag}(\mathbf{a})^{-1}\boldsymbol{\lambda}) \\ &= \mathbb{E}(\|\boldsymbol{\theta}^{m\top}\operatorname{diag}(\mathbf{a})^{-1}\boldsymbol{\lambda}\|^2), \end{aligned}$$

which implies that $\boldsymbol{\theta}^{m\top}\operatorname{diag}(\mathbf{a})^{-1}\boldsymbol{\lambda} \overset{a.s.}{=} 0$. Since $\boldsymbol{\lambda} \neq \mathbf{0}$, we can assume, without loss of generality, that the first t elements of $\boldsymbol{\lambda}$, $\lambda_1, \ldots, \lambda_t > 0$, the next s elements of λ, $\lambda_{t+1}, \ldots, \lambda_{t+s} < 0$, and the remaining elements are 0 for some $s, t : s > 0, t > 0, s + t \leq K$. Therefore, if we define two nonnegative and nonzero row vectors $\mathbf{b}_1 := b\left[\lambda_1 a_1^{-1}, \ldots, \lambda_t a_t^{-1} 0, \ldots, 0\right]$ and $\mathbf{b}_2 := -b\left[0, \ldots, 0, \lambda_{t+1}a_{t+1}^{-1}, \ldots, \lambda_s a_s^{-1}, 0, \ldots, 0\right]$, where $b > 0$ is a constant, then $\mathbf{b}_1\boldsymbol{\theta}^m \overset{a.s.}{=} \mathbf{b}_2\boldsymbol{\theta}^m$.

Now we construct two topic matrices $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ as follows. Let $\mathbf{b}_1$ be the first row and $\mathbf{b}_2$ the second in $\boldsymbol{\beta}_1$. Let $\mathbf{b}_2$ be the first row and $\mathbf{b}_1$ the second in $\boldsymbol{\beta}_2$. Let $\mathbf{B} \in \mathbb{R}^{W-2\times K}$ be a valid topic matrix and assume that it is **separable**. Set the remaining $(W-2)$ rows of both $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ to be $\mathbf{B}(I_K - \operatorname{diag}(\mathbf{b}_1 + \mathbf{b}_2))$. We can choose b to be small enough to ensure that each element of $(\mathbf{b}_1 + \mathbf{b}_2)$ is strictly less than 1. This will ensure that $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ are column-stochastic and therefore valid topic matrices. We note that the supports of $\mathbf{b}_1$ and $\mathbf{b}_2$ are disjoint and both are non-empty. They appear in distinct topics.

By construction, $\boldsymbol{\beta}^{(1)}\boldsymbol{\theta} \overset{a.s.}{=} \boldsymbol{\beta}^{(2)}\boldsymbol{\theta} \Rightarrow$ the distribution of the observation $\mathbf{X}$ conditioned on $\boldsymbol{\theta}$ is the same for both models. Marginalizing over $\boldsymbol{\theta}$, the distributions of $\mathbf{X}$ under the topic matrices are the same. Thus no algorithm can distinguish between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ based on $\mathbf{X}$. □

C. Proof of Proposition 1 and Proposition 2

Proposition 1 and Proposition 2 summarizes the relationships between the full-rank, affine-independence, simplicial, and diagonal-dominance conditions. Here we consider all the pairwise implication separately.

(1) $\bar{\mathbf{R}}$ is γ_a-affine-independent $\Rightarrow$ $\bar{\mathbf{R}}$ is at least γ_a-simplicial.

Proof. By definition of affine independence, $\|\sum_{k=1}^K \lambda_k\bar{\mathbf{R}}_k\|_2 \geq \gamma_a\|\boldsymbol{\lambda}\|_2 > 0$ for all $\boldsymbol{\lambda} \in \mathbb{R}^K$ such that $\sum_{k=1}^K \lambda_k = 0$ and $\boldsymbol{\lambda} \neq \mathbf{0}$. If for each $i \in [K]$ we set $\lambda_k = 1$ for $k = i$ and choose $\lambda_k \leq 0,\ \forall k \neq i$ then (i) $\|\boldsymbol{\lambda}\|_2 \geq 1$, (ii) $\{-\lambda_k, k \neq i\}$ are convex weights, i.e., they are nonnegative and sum to 1, and (iii) $\sum_{k=1}^K \lambda_k\bar{\mathbf{R}}_k = \bar{\mathbf{R}}_i - \sum_{k\neq i}(-\lambda_k)\bar{\mathbf{R}}_k$. Therefore, for all $i \in [K]$, $\|\bar{\mathbf{R}}_i - \sum_{k\neq i}(-\lambda_k)\bar{\mathbf{R}}_k\|_2 \geq \gamma_a > 0$ which proves that $\bar{\mathbf{R}}$ is at least γ_a-simplicial.

For the reverse implication, consider

$$\bar{\mathbf{R}} = \begin{bmatrix} 1 & 0 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0 \\ 0.5 & 0.5 & 0 & 1 \end{bmatrix}.$$

It is simplicial but is not affine independent (the $1, 1, -1, -1$ combination of the 4 rows would be $\mathbf{0}$). □

(2) $\bar{\mathbf{R}}$ is full rank with minimum eigenvalue $\gamma_r \Rightarrow \bar{\mathbf{R}}$ is at least γ_r-affine-independent.

Proof. The Rayleigh-quotient characterization of the minimum eigenvalue of a symmetric, positive-definite matrix $\bar{\mathbf{R}}$ gives $\min_{\boldsymbol{\lambda}\neq\mathbf{0}} \|\boldsymbol{\lambda}^\top\mathbf{R}\|_2/\|\boldsymbol{\lambda}\|_2 = \gamma_r > 0$. Therefore, $\min_{\boldsymbol{\lambda}\neq\mathbf{0},\mathbf{1}^\top\boldsymbol{\lambda}=0} \|\boldsymbol{\lambda}^\top\mathbf{R}\|_2/\|\boldsymbol{\lambda}\|_2 \geq \gamma_r > 0$. One can construct examples that contradict the reverse implication:

$$\bar{\mathbf{R}} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

which is affine independent, but not linear independent. □

(3) $\bar{\mathbf{R}}$ is γ_d-diagonal-dominant $\Rightarrow$ $\bar{\mathbf{R}}$ is at least γ_d-simplicial.

Proof. Noting that $\bar{\mathbf{R}}_{i,i} - \bar{\mathbf{R}}_{i,j} \geq \gamma_d > 0$ for all i, j, then the distance of the first row of $\bar{\mathbf{R}}$, $\bar{\mathbf{R}}_1$, to any convex combination of the remaining rows, $\sum_{j=2}^K c_j\bar{\mathbf{R}}_j$, where $c_2, \ldots, c_K$ are convex combination weights, can be lower bounded by, $\|\bar{\mathbf{R}}_1 - \sum_{j=2}^K c_j\bar{\mathbf{R}}_j\|_2 \geq |\bar{\mathbf{R}}_{1,1} - \sum_{j=2}^K c_j\bar{\mathbf{R}}_{j,1}| = |\sum_{j=2}^K c_j(\bar{\mathbf{R}}_{1,1} - \bar{\mathbf{R}}_{j,1})| \geq \gamma_d > 0$. Therefore, $\bar{\mathbf{R}}$ is at least γ_d-simplicial. It is straightforward to construct examples that contradict the reverse implication:

$$\bar{\mathbf{R}} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

which is affine independent, hence simplicial, but not diagonal-dominant. □

(4) $\bar{\mathbf{R}}$ being diagonal-dominant neither implies nor is implied by $\bar{\mathbf{R}}$ being affine-independent.

Proof. Consider the following two examples:

$$\bar{\mathbf{R}} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

and

$$\bar{\mathbf{R}} = \begin{bmatrix} 1 & 0 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0 \\ 0.5 & 0.5 & 0 & 1 \end{bmatrix}.$$

They are the examples for the two sides of this assertion. □

D. Proof of Lemma 3

Proof. Recall that$\bar{\mathbf{A}} = \bar{\boldsymbol{\beta}}\bar{\boldsymbol{\theta}}$ where $\bar{\mathbf{A}}$ and $\bar{\boldsymbol{\theta}}$ are row-normalized version of $\mathbf{A}$ and $\boldsymbol{\theta}$, $\bar{\boldsymbol{\beta}} := \mathrm{diag}(\mathbf{A}\mathbf{1})^{-1}\boldsymbol{\beta}\,\mathrm{diag}(\boldsymbol{\theta}\mathbf{1})$. $\bar{\boldsymbol{\beta}}$ is row-stochastic and is separable if $\boldsymbol{\beta}$ is separable. If w is a novel word of topic k, $\bar{\beta}_{wk} = 1$ and $\bar{\beta}_{wj} = 0,\ \forall j \neq k$. We have then $\bar{\mathbf{A}}_w = \bar{\boldsymbol{\theta}}_k$. If w is a non-novel word, $\bar{\mathbf{A}}_w = \sum_k \bar{\beta}_{wk}\bar{\boldsymbol{\theta}}_k$ is a convex combination of the rows of $\bar{\boldsymbol{\theta}}$.

We next prove that if $\bar{\mathbf{R}}$ is γ_s-simplicial with some constant $\gamma_s > 0$, then, the random matrix $\bar{\boldsymbol{\theta}}$ is also simplicial with high probability, i.e., for any $\mathbf{c} \in \mathbb{R}^K$ such that $c_k = 1, c_j \leq 0, j \neq k, \sum_{j\neq k} -c_j = 1, k \in [K]$, the M-dimensional vector $\mathbf{c}^\top\bar{\boldsymbol{\theta}}$ is not all-zero with high probability. In another words, we need to show that the maximum absolute value of the M entries in $\mathbf{c}^\top\bar{\boldsymbol{\theta}}$ is strictly positive. Noting that the m-th entry of $\mathbf{c}^\top\bar{\boldsymbol{\theta}}$ (scaled by M) is

$$\begin{aligned} M\mathbf{c}^\top\bar{\boldsymbol{\theta}}^m =& \mathbf{c}^\top\,\mathrm{diag}(\mathbf{a})^{-1}\boldsymbol{\theta}^m \\ &+ \mathbf{c}^\top(\mathrm{diag}(\sum_d \boldsymbol{\theta}^d/M)^{-1} - \mathrm{diag}(\mathbf{a})^{-1})\boldsymbol{\theta}^m \end{aligned}$$

the absolute value can be lower bounded as follows,

$$\begin{aligned} |M\mathbf{c}^\top\bar{\boldsymbol{\theta}}^m| \geq& |\mathbf{c}^\top\,\mathrm{diag}(\mathbf{a})^{-1}\boldsymbol{\theta}^m| \\ &- |\mathbf{c}^\top(\mathrm{diag}(\sum_d \boldsymbol{\theta}^d/M)^{-1} - \mathrm{diag}(\mathbf{a})^{-1})\boldsymbol{\theta}^m| \end{aligned} \quad (9)$$

The key ideas are: (i) as M increases, the second term in Eq. (9) converges to 0, and (ii) the maximum of the first term in Eq. (9) among $m = 1, \ldots, M$ is strictly above zero with high probability. For (i), recall that $\mathbf{a} = \mathbb{E}(\boldsymbol{\theta}^m)$ and $0 \leq \theta_k^m \leq 1$, by Hoeffding's lemma $\forall t > 0$,

$$\Pr(\|\sum_d \boldsymbol{\theta}^d/M - \mathbf{a}\|_\infty \geq t) \leq 2K\exp(-2Mt^2)$$

Also note that $\forall 0 < \epsilon < 1$,

$$\begin{aligned} &\|\sum_d \boldsymbol{\theta}^d/M - \mathbf{a}\|_\infty \leq \epsilon a_{\min}^2/2 \\ \Rightarrow& \|(\mathrm{diag}(\sum_d \boldsymbol{\theta}^d/M)^{-1} - \mathrm{diag}(\mathbf{a})^{-1})\|_\infty \leq \epsilon \\ \Rightarrow& |\mathbf{c}^\top(\mathrm{diag}(\sum_d \boldsymbol{\theta}^d/M)^{-1} - \mathrm{diag}(\mathbf{a})^{-1})\boldsymbol{\theta}^m| \leq \epsilon \end{aligned}$$

where $a_{\min}$ is the minimum entry of $\mathbf{a}$. The last inequality is true since $\sum_{k=1}^K \theta_k^m = 1$. In sum, we have

$$\begin{aligned} &\Pr(|\mathbf{c}^\top(\mathrm{diag}(\sum_d \boldsymbol{\theta}^d/M)^{-1} - \mathrm{diag}(\mathbf{a})^{-1})\boldsymbol{\theta}^m| > \epsilon) \\ &\leq 2K\exp(-M\epsilon^2 a_{\min}^4/2) \end{aligned} \quad (10)$$

For (ii), recall that $\bar{\mathbf{R}}$ is γ_s-simplicial and $\|\mathbf{c}^\top\bar{\mathbf{R}}\| \geq \gamma_s$. Therefore, $\mathbf{c}^\top\bar{\mathbf{R}}\mathbf{c} = \mathbf{c}^\top\bar{\mathbf{R}}\bar{\mathbf{R}}^\dagger\bar{\mathbf{R}}\mathbf{c} \geq \frac{\gamma_s^2}{\lambda_{\mathsf{max}}}$ where λ_{max} is the maximum singular value of $\bar{\mathbf{R}}$. Noting that $\bar{\mathbf{R}} = \mathrm{diag}(\mathbf{a})^{-1}\,\mathbb{E}(\boldsymbol{\theta}^m\boldsymbol{\theta}^{m\top})\,\mathrm{diag}(\mathbf{a})^{-1}$, we get

$$\mathbb{E}(|\mathbf{c}^\top\,\mathrm{diag}(\mathbf{a})^{-1}\boldsymbol{\theta}^m|^2) \geq \frac{\gamma_s^2}{\lambda_{\mathsf{max}}} \quad (11)$$

For convenience, let $x_m := |\mathbf{c}^\top\,\mathrm{diag}(\mathbf{a})^{-1}\boldsymbol{\theta}^m|^2 \leq 1/a_{\mathtt{min}}^2$. Then, by Hoeffding's lemma,

$$\Pr(\mathbb{E}(x_m) - \sum_{m=1}^M x_m/M \geq \frac{\gamma_s^2}{2\lambda_{\mathsf{max}}}) \leq \exp(-M\gamma_s^4 a_{\mathtt{min}}^4/2\lambda_{\mathsf{max}}^2)$$

Combining Eq. (11) we get

$$\Pr(\sum_{m=1}^M x_m/M \leq \frac{\gamma_s^2}{2\lambda_{\mathsf{max}}}) \leq \exp(-M\gamma_s^4 a_{\mathtt{min}}^4/2\lambda_{\mathsf{max}}^2)$$

Hence

$$\Pr(\max_{m=1}^M x_m \leq \frac{\gamma_s^2}{2\lambda_{\mathsf{max}}}) \leq \exp(-M\gamma_s^4 a_{\mathtt{min}}^4/2\lambda_{\mathsf{max}}^2) \quad (12)$$

i.e., the maximum absolute value of the first term in Eq. (9) is greater than $\gamma_s/\sqrt{2\lambda_{\mathsf{max}}}$ with high probability.

To sum up, if we set $\epsilon = \gamma_s/\sqrt{2\lambda_{\mathsf{max}}}$ in Eq. (10), we get

$$\begin{aligned} \Pr(\max_{m=1}^M |\mathbf{c}^\top\bar{\boldsymbol{\theta}}^m| = 0) \leq& \Pr(\max_{m=1}^M x_m \leq \frac{\gamma_s^2}{2\lambda_{\mathsf{max}}}) \\ &+ \Pr(|\mathbf{c}^\top(\mathrm{diag}(\frac{\mathbf{a}}{\sum_d \boldsymbol{\theta}_d}) - I)\boldsymbol{\theta}^m| > \epsilon) \\ \leq& \exp(-M\gamma_s^4 a_{\mathtt{min}}^4/2\lambda_{\mathsf{max}}^2) \\ &+ 2K\exp(-M\gamma_s^2 a_{\min}^4/4\lambda_{max}) \end{aligned}$$

To summarize, the probability that $\bar{\theta}$ is not simplicial is at most $\exp(-M\gamma_s^4 a_{\mathtt{min}}^4/2\lambda_{\mathsf{max}}^2) + 2K\exp(-M\gamma_s^2 a_{\min}^4/4\lambda_{max})$. This converges to 0 exponentially fast as $M \to \infty$. Therefore, with high probability, all the row-vectors of $\bar{\boldsymbol{\theta}}$ are extreme points of the convex hull they form and this concludes our proof. □

E. Proof of Lemma 4

Proof. We first show that if $\bar{\mathbf{R}}$ is γ_a affine-independent, $\bar{\boldsymbol{\theta}}$ is also affine-independent with high probability, i.e., $\forall \mathbf{c} \in \mathbb{R}^K$ such that $\mathbf{c} \neq \mathbf{0}, \sum_k c_k = 0$, $\mathbf{c}^\top\bar{\boldsymbol{\theta}}$ is not all-zero vector with high probability. Our proof is similar to that of Lemma 3. We first re-write the m-th entry of $\mathbf{c}^\top\bar{\boldsymbol{\theta}}$ (with some scaling) as,

$$\begin{aligned} M\mathbf{c}^\top\bar{\boldsymbol{\theta}}^m =& \mathbf{c}^\top\,\mathrm{diag}(\mathbf{a})^{-1}\boldsymbol{\theta}^m \\ &+ \mathbf{c}^\top(\mathrm{diag}(\sum_d \boldsymbol{\theta}^d/M)^{-1} - \mathrm{diag}(\mathbf{a})^{-1})\boldsymbol{\theta}^m \end{aligned}$$

and lower bound its absolute value by

$$\begin{aligned} |M\mathbf{c}^\top\bar{\boldsymbol{\theta}}^m| \geq& |\mathbf{c}^\top\,\mathrm{diag}(\mathbf{a})^{-1}\boldsymbol{\theta}^m| \\ &- |\mathbf{c}^\top(\mathrm{diag}(\sum_d \boldsymbol{\theta}^d/M)^{-1} - \mathrm{diag}(\mathbf{a})^{-1})\boldsymbol{\theta}^m| \end{aligned} \quad (13)$$

We will then show that: (i) as M increases, the second term in Eq. (13) converges to 0, and (ii) the maximum of the first term in Eq. (13) among M iid samples is strictly above zero with high probability. For (i), by the Cauchy-Schwartz inequality

$$\begin{aligned} &|\mathbf{c}^\top(\mathrm{diag}(\sum_d \boldsymbol{\theta}^d/M)^{-1} - \mathrm{diag}(\mathbf{a})^{-1})\boldsymbol{\theta}^m| \\ \leq& \|\mathbf{c}\|_2\|(\mathrm{diag}(\sum_d \boldsymbol{\theta}^d/M)^{-1} - \mathrm{diag}(\mathbf{a})^{-1})\boldsymbol{\theta}^m\|_2 \\ \leq& \|\mathbf{c}\|_2\|(\mathrm{diag}(\sum_d \boldsymbol{\theta}^d/M)^{-1} - \mathrm{diag}(\mathbf{a})^{-1})\|_\infty \end{aligned}$$

Here the last inequality is true since $\theta_k^m \leq 1, \sum_k \theta_k^m = 1$. Similar to Eq. (10), we have,

$$\Pr(|(\mathrm{diag}(\sum_d \boldsymbol{\theta}^d/M)^{-1} - \mathrm{diag}(\mathbf{a})^{-1})\boldsymbol{\theta}^m| \geq \|\mathbf{c}\|_2\epsilon)$$

$$\leq 2K \exp(-M\epsilon^2 a_{\min}^4/4) \quad (14)$$

for any $0 < \epsilon < 1$, $a_{\min}$ is the minimum entry of $\mathbf{a}$. For (ii), recall that by definition, $\|\mathbf{c}^\top \bar{\mathbf{R}}\|_2 \geq \gamma_a \|\mathbf{c}\|_2$. Hence $\mathbf{c}^\top \bar{\mathbf{R}} \mathbf{c} \geq \gamma_a^2 \|\mathbf{c}\|_2^2/\lambda_{\max}$. Therefore, by the construction of $\bar{\mathbf{R}}$, we have,

$$\mathbb{E}(|\mathbf{c}^\top \operatorname{diag}(\mathbf{a})^{-1}\boldsymbol{\theta}^m|^2/\|\mathbf{c}\|_2^2) \geq \frac{\gamma_a^2}{\lambda_{\max}} \quad (15)$$

For convenience, let $x_m := |\mathbf{c}^\top \operatorname{diag}(\mathbf{a})^{-1}\boldsymbol{\theta}^m|^2/\|\mathbf{c}\|_2^2 \leq 1/a_{\min}^2$. Following the same procedure as in Eq. (12), we have,

$$\Pr(\max_{m=1}^{M} x_m \leq \frac{\gamma_a^2}{2\lambda_{\max}}) \leq \exp(-M\gamma_a^4 a_{\min}^4/2\lambda_{\max}^2) \quad (16)$$

Therefore, if we set in Eq. (14) $\epsilon = \gamma/\sqrt{2\lambda_{\max}}$, we get,

$$\begin{aligned}\Pr(\max_{m=1}^{M} |\mathbf{c}^\top \bar{\boldsymbol{\theta}}^m| \leq 0) \leq & \exp(-M\gamma_a^4 a_{\min}^4/2\lambda_{\max}^2) \\ & + 2K\exp(-M\gamma_a^2 a_{\min}^4/4\lambda_{\max})\end{aligned}$$

In summary, if $\bar{\mathbf{R}}$ is γ_a affine-independent, $\bar{\boldsymbol{\theta}}$ is also affine-independent with high probability.

Now we turn to prove Lemma 4. By Lemma 3, detecting K distinct novel words for K topics is equivalent to knowing $\bar{\boldsymbol{\theta}}$ up to a row permutation. Noting that $\bar{\mathbf{A}}_w = \sum_k \bar{\beta}_{wk}\bar{\boldsymbol{\theta}}_k$. it follows that $\bar{\beta}_{wk}, k = 1, \ldots, K$ is one optimal solution to the following constrained optimization problem:

$$\min \quad \|\bar{\mathbf{A}}_w - \sum_{k=1}^{K} b_k \bar{\boldsymbol{\theta}}_k\|^2 \text{ s.t } \quad b_k \geq 0, \sum_{k=1}^{K} b_k = 1$$

Since $\bar{\boldsymbol{\theta}}$ is affine-independent with high probability, therefore, this optimal solution is unique with high probability. If this is not true, then there would exist two distinct solutions $b_1^1, \ldots, b_K^1$ and $b_1^2, \ldots, b_K^2$ such that $\bar{\mathbf{A}}_w = \sum_{k=1}^K b_k^1 \bar{\boldsymbol{\theta}}_k = \sum_{k=1}^K b_k^2 \bar{\boldsymbol{\theta}}_k$. $\sum b_k^1 = \sum b_k^2 = 1$. We would then obtain

$$\sum_{k=1}^{K} (b_k^1 - b_k^2)\bar{\boldsymbol{\theta}}_k = \mathbf{0}$$

where the coefficients $b_k^1 - b_k^2$ are not all zero and $\sum_k b_k^1 - b_k^2 = 0$. This would contradict the affine-independence definition.

Finally, we check the renormalization steps. Recall that since $\operatorname{diag}(\mathbf{A}\mathbf{1})\bar{\boldsymbol{\beta}} = \boldsymbol{\beta}\operatorname{diag}(\boldsymbol{\theta}\mathbf{1})$, $\operatorname{diag}(\mathbf{A}\mathbf{1})$ can be directly obtained from the observations. So we can first renormalize the rows of $\bar{\boldsymbol{\beta}}$. Removing $\operatorname{diag}(\boldsymbol{\theta}\mathbf{1})$ is then simply a column renormalization operation (recall that $\boldsymbol{\beta}$ is column-stochastic). It is not necessary to know the exact the value of $\operatorname{diag}(\boldsymbol{\theta}\mathbf{1})$.

To sum up, by solving a constrained linear regression followed by suitable row renormalization, we can obtain a unique solution which is the ground truth topic matrix. This concludes the proof of Lemma 4. □

F. Proof of Lemma 5

Lemma 5 establishes the second order co-occurrence estimator in Eq. (1). We first provide a generic method to establish the explicit convergence bound for a function $\psi(\mathbf{X})$ of d random variables $X_1, \ldots, X_d$, then apply it to establish Lemma 5

Proposition 3. *Let $\mathbf{X} = [X_1, \ldots, X_d]$ be d random variables and $\mathbf{a} = [a_1, \ldots, a_d]$ be positive constants. Let $\mathcal{E} := \bigcup_{i\in\mathcal{I}} \{|X_i - a_i| \geq \delta_i\}$ for some constants $\delta_i > 0$, and $\psi(\mathbf{X})$ be a continuously differentiable function in $\mathcal{C} := \mathcal{E}^c$. If for $i = 1, \ldots, d$, $\Pr(|X_i - a_i| \geq \epsilon) \leq f_i(\epsilon)$ are the individual convergence rates and $\max_{X\in\mathcal{C}} |\partial_i \psi(\mathbf{X})| \leq C_i$, then,*

$$\Pr(|\psi(\mathbf{X}) - \psi(\mathbf{a})| \geq \epsilon) \leq \sum_i f_i(\delta_i) + \sum_{i=1} f_i(\frac{\epsilon}{dC_i})$$

Proof. Since $\psi(\mathbf{X})$ is continuously differentiable in $\mathcal{C}$, $\forall \mathbf{X} \in \mathbf{C}, \exists \lambda \in (0,1)$ such that

$$\psi(\mathbf{X}) - \psi(\mathbf{a}) = \nabla^\top \psi((1-\lambda)\mathbf{a} + \lambda\mathbf{X}) \cdot (\mathbf{X} - \mathbf{a})$$

Therefore,

$$\begin{aligned}&\Pr(|\psi(\mathbf{X}) - \psi(\mathbf{a})| \geq \epsilon) \\ \leq & \Pr(\mathbf{X} \in \mathcal{E}) + \\ & \Pr(\sum_{i=1}^{d} |\partial_i \psi((1-\lambda)\mathbf{a} + \lambda\mathbf{X})||X_i - a_i| \geq \epsilon | \mathbf{X} \in \mathcal{C}) \\ \leq & \sum_{i\in\mathcal{I}} \Pr(|X_i - a_i| \geq \delta_i) + \\ & \sum_{i=1}^{d} \Pr(\max_{\mathbf{x}\in\mathcal{C}} |\partial_i \psi(\mathbf{x})||X_i - a_i| \geq \epsilon/d) \\ = & \sum_{i\in\mathcal{I}} f_i(\delta_i) + \sum_{i=1} f_i(\frac{\epsilon}{dC_i})\end{aligned}$$

□

Now we turn to prove Lemma 5. Recall that $\bar{\mathbf{X}}$ and $\bar{\mathbf{X}}'$ are obtained from $\mathbf{X}$ by first splitting each user's comparisons into two independent halves and then rescaling the rows to make them row-stochastic hence $\bar{\mathbf{X}} = \operatorname{diag}^{-1}(\mathbf{X}\mathbf{1})\mathbf{X}$. Also recall that $\bar{\boldsymbol{\beta}} = \operatorname{diag}^{-1}(\boldsymbol{\beta}\mathbf{a})\boldsymbol{\beta}\operatorname{diag}(\mathbf{a})$, $\bar{\mathbf{R}} = \operatorname{diag}^{-1}(\mathbf{a})\mathbf{R}\operatorname{diag}^{-1}(\mathbf{a})$, and $\bar{\boldsymbol{\beta}}$ is row stochastic. For any $1 \leq i, j \leq W$,

$$\begin{aligned}\widehat{E}_{i,j} &= M\frac{1}{\sum_{m=1}^{M} X'_{i,m}}(\sum_{m=1}^{M} X'_{i,m}X_{j,m})\frac{1}{\sum_{m=1}^{M} X_{i,m}} \\ &= \frac{1/M \sum_{m=1}^{M}(X'_{i,m}X_{j,m})}{(1/M \sum_{m=1}^{M} X'_{i,m})(1/M \sum_{m=1}^{M} X_{j,m})} \\ &= \frac{\frac{1}{MN^2}\sum_{m=1,n=1,n'=1}^{M,N,N} \mathbb{I}(w_{m,n} = i)\mathbb{I}(w'_{m,n'} = j)}{\frac{1}{MN}\sum_{m=1,n=1}^{M,N} \mathbb{I}(w_{m,n} = i)\frac{1}{MN}\sum_{m=1,n=1}^{M,N} \mathbb{I}(w'_{m,n} = i)} \\ &:= \frac{F_{i,j}(M,N)}{G_i(M,N)H_j(M,N)}\end{aligned}$$

From the Strong Law of Large Numbers and the generative topic modeling procedure,

$$\begin{aligned}F_{i,j}(M,N) &\xrightarrow{a.s.} \mathbb{E}(\mathbb{I}(w_{m,n} = i)\mathbb{I}(w'_{m,n'} = j)) \\ &= (\boldsymbol{\beta}\mathbf{R}\boldsymbol{\beta}^\top)_{i,j} := p_{i,j}\end{aligned}$$

$$G_i(M,N) \xrightarrow{a.s.} \mathbb{E}(\mathbb{I}(w'_{m,n}=i)) = (\boldsymbol{\beta}\mathbf{a})_i := p_i$$
$$H_i(M,N) \xrightarrow{a.s.} \mathbb{E}(\mathbb{I}(w_{m,n}=j)) = (\boldsymbol{\beta}\mathbf{a})_j := p_j$$

and $\frac{(\boldsymbol{\beta}\mathbf{R}\boldsymbol{\beta}^\top)_{i,j}}{(\boldsymbol{\beta}\mathbf{a})_i(\boldsymbol{\beta}\mathbf{a})_j} = \mathbf{E}_{i,j}$ by definition. Using McDiarmid's inequality, we obtain

$$\Pr(|F_{i,j}-p_{i,j}| \geq \epsilon) \leq 2\exp(-\epsilon^2 MN)$$
$$\Pr(|G_i-p_i| \geq \epsilon) \leq 2\exp(-2\epsilon^2 MN)$$
$$\Pr(|H_j-p_j| \geq \epsilon) \leq 2\exp(-2\epsilon^2 MN)$$

In order to calculate $\Pr\{|\frac{F_{i,j}}{G_iH_j} - \frac{p_{i,j}}{p_ip_j}| \geq \epsilon\}$, we apply the results from Proposition 3. Let $\psi(x_1,x_2,x_3) = \frac{x_1}{x_2x_3}$ with $x_1,x_2,x_3 > 0$, and $a_1 = p_{i,j}$, $a_2 = p_i$, $a_3 = p_j$. Let $\mathcal{I} = \{2,3\}$, $\delta_2 = \gamma p_i$, and $\delta_3 = \gamma p_j$. Then $|\partial_1\psi| = \frac{1}{x_2x_3}$, $|\partial_2\psi| = \frac{x_1}{x_2^2x_3}$, and $|\partial_3\psi| = \frac{x_1}{x_2x_3^2}$. If $F_{i,j}=x_1$, $G_i = x_2$, and $H_j = x_3$, then $F_{i,j} \leq G_i$, $F_{i,j} \leq H_j$. Then note that

$$C_1 = \max_{\mathcal{C}} |\partial_1\psi| = \max_{\mathcal{C}} \frac{1}{G_iH_j} \leq \frac{1}{(1-\gamma)^2p_ip_j}$$
$$C_2 = \max_{\mathcal{C}} |\partial_2\psi| = \max_{\mathcal{C}} \frac{F_{i,j}}{G_i^2H_j} \leq \max_{\mathcal{C}} \frac{1}{G_iH_j} \leq \frac{1}{(1-\gamma)^2p_ip_j}$$
$$C_3 = \max_{\mathcal{C}} |\partial_3\psi| = \max_{\mathcal{C}} \frac{F_{i,j}}{G_iH_j^2} \leq \max_{\mathcal{C}} \frac{1}{G_iH_j} \leq \frac{1}{(1-\gamma)^2p_ip_j}$$

By applying Proposition 3, we get

$$\begin{aligned}
&\Pr\{|\frac{F_{i,j}}{G_iH_j} - \frac{p_{i,j}}{p_ip_j}| \geq \epsilon\} \\
\leq &\exp(-2\gamma^2p_i^2MN) + \exp(-2\gamma^2p_j^2MN) \\
&+ 2\exp(-\epsilon^2(1-\gamma)^4(p_ip_j)^2MN/9) \\
&+ 4\exp(-2\epsilon^2(1-\gamma)^4(p_ip_j)^2MN/9) \\
\leq &2\exp(-2\gamma^2\eta^2MN) + 6\exp(-\epsilon^2(1-\gamma)^4\eta^4MN/9)
\end{aligned}$$

where $\eta = \min_{1\leq i\leq W} p_i$. There are many strategies for optimizing the free parameter γ. We set $2\gamma^2 = \frac{(1-\gamma)^4}{9}$ and solve for γ to obtain

$$\Pr\{|\frac{F_{i,j}}{G_iH_j} - \frac{p_{i,j}}{p_ip_j}| \geq \epsilon\} \leq 8\exp(-\epsilon^2\eta^4MN/20)$$

Finally, by applying the union bound to the W^2 entries in $\widehat{\mathbf{E}}$, we obtain the claimed result.

G. Proof of Lemma 6

Proof. We first show that when $\bar{\mathbf{R}}$ is γ_s simplicial and $\boldsymbol{\beta}$ is separable, then $\mathbf{Y} = \bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top$ is at least γ_s-simplicial. Without loss of generality we assume that word $1,\ldots,K$ are the novel words for topic 1 to K. By definition, $\bar{\boldsymbol{\beta}}^\top = [\mathbf{I}_K, \mathbf{B}]$ hence $\mathbf{Y} = \bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top = [\bar{\mathbf{R}}, \bar{\mathbf{R}}\mathbf{B}]$. Therefore, for convex combination weights $c_2,\ldots,c_K \geq 0$ such that $\sum_{j=2}^K c_j = 1$,

$$\|\mathbf{Y}_1 - \sum_{j=2}^K c_j\mathbf{Y}_j\| \geq \|\bar{\mathbf{R}}_1 - \sum_{j=2}^K c_j\bar{\mathbf{R}}_j\| \geq \gamma_s > 0$$

Therefore the first row vector $\mathbf{Y}_1$ is at least γ_s distant away from the convex hull of the remaining rows. Similarly, any row of $\mathbf{Y}$ is at least γ_s distant away from the convex hull of the remaining rows hence $\mathbf{Y}$ is at least γ_s simplicial. The rest of the proof will be exactly the same as for Lemma 6. □

H. Proof of Lemma 7

Proof. We first show that when $\bar{\mathbf{R}}$ is γ_a affine independent and $\boldsymbol{\beta}$ is separable, then $\mathbf{Y} = \bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top$ is at least γ_a affine independent. Similarly as in the proof of Lemma 6, we assume that word $1,\ldots,K$ are the novel words for topic 1 to K. By definition, $\bar{\boldsymbol{\beta}}^\top = [\mathbf{I}_K, \mathbf{B}]$ hence $\mathbf{Y} = \bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top = [\bar{\mathbf{R}}, \bar{\mathbf{R}}\mathbf{B}]$. $\forall\boldsymbol{\lambda} \in \mathbb{R}^K$ such that $\boldsymbol{\lambda} \neq \mathbf{0}$, $\sum_{k=1}^K \lambda_k = 0$, then,

$$\|\sum_{k=1}^K \mathbf{Y}_k\|_2/\|\boldsymbol{\lambda}\|_2 \geq \|\sum_{k=1}^K \bar{\mathbf{R}}_k\|_2/\|\boldsymbol{\lambda}\|_2 \geq \gamma_a$$

Hence $\mathbf{Y}$ is affine independent. The The rest of the proof will be exactly the same as that for Lemma 4.

We note that once the novel words for K topics are detection, we can use only the corresponding columns of $\mathbf{E}$ for linear regression. Formally, let $\mathbf{E}^*$ be the $W \times K$ matrix formed by the columns of the $\mathbf{E}$ that correspond to K distinct novel words. Then, $\mathbf{E}^* = \bar{\boldsymbol{\beta}}\bar{\mathbf{R}}$. The rest of the proof is again the same as that for Lemma 4. □

I. Proof of Lemma 8

Proof. We first check that if $q_w > 0$, w must be a novel word. Without loss of generality let word $1,\ldots,K$ be novel words for K distinct topics. $\forall w$, $\mathbf{E}_w = \sum \bar{\beta}_{wk}\mathbf{E}_k$. $\forall\mathbf{d} \in \mathbb{R}^W$,

$$\langle\mathbf{E}_w, \mathbf{d}\rangle = \sum \bar{\beta}_{wk}\langle\mathbf{E}_k, \mathbf{d}\rangle \leq \max_k\langle\mathbf{E}_k, \mathbf{d}\rangle$$

and the last equality holds if, and only if, there exist some k such that $\bar{\beta}_{wk} = 1$ which implies w is a novel words.

We then show that for a novel word w, $q_w > 0$. We need to show for each topic k, when $\mathbf{d}$ is sampled from an isotropic distribution in $\mathbf{R}^W$, there exist a set of directions $\mathbf{d}$ with nonzero probability such that $\langle\mathbf{E}_k, \mathbf{d}\rangle > \langle\mathbf{E}_l, \mathbf{d}\rangle$ for $l = 1,\ldots,K, l \neq k$. First, one can check by definition that $\mathbf{Y} = (\mathbf{E}_1^\top,\ldots,\mathbf{E}_K^\top)^\top = \bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top$ is at least γ_s-simplicial if $\bar{\mathbf{R}}$ is γ_s-simplicial. Let $\mathbf{E}_1^*$ be the projection of $\mathbf{E}_1$ onto the simplex formed by the remaining row vectors $\mathbf{E}_2,\ldots,\mathbf{E}_K$. By the orthogonality principle, $\langle\mathbf{E}_1 - \mathbf{E}_1^*, \mathbf{E}_k - \mathbf{E}_1^*\rangle \leq 0$ for $k = 2,\ldots,K$. Therefore, for $\mathbf{d}^1 = \mathbf{E}_1^\top - \mathbf{E}_1^{*\top}$,

$$\mathbf{E}_1\mathbf{d}^1 - \mathbf{E}_k\mathbf{d}^1 = \|\mathbf{d}^1\|^2 - (\mathbf{E}_k - \mathbf{E}_1^*)\mathbf{d}^1 \geq \gamma_s^2 > 0$$

Due to the continuity of the inner product, there exist a neighbor on the unite sphere around $\mathbf{d}^1/\|\mathbf{d}^1\|_2$ that $\mathbf{E}_1$ has maximum projection value. This conclude our proof. □

J. Proof of Theorem 2

Proof. We first consider the random projection steps (step 3 to 12 in Alg. 2). For projection along direction $\mathbf{d}^r$, we first calculate projection values $\mathbf{r} = \bar{\mathbf{X}}'\bar{\mathbf{X}}^\top\mathbf{d}^r$, find the maximizer index i^* and the corresponding set $\hat{J}_{i^*}$, and then evaluate $\mathbb{I}(\forall j \in \hat{J}_w, v_w > v_j)$ for all the words w in $\hat{J}_{i^*}^c = \{1,\ldots,W\} \setminus \hat{J}_{i^*}$. (I) The set $\hat{J}_{i^*}^c$ have up to $|\mathcal{C}_k|$ elements asymptotically, where k is the topic associated with word i^*. This is considered a small constant $\mathcal{O}(1)$; (II) Note that $\widehat{\mathbf{E}}\mathbf{d}_r = M\bar{\mathbf{X}}'(\bar{\mathbf{X}}^\top\mathbf{d}_r)$ and each column of $\bar{\mathbf{X}}$ has at most $N \ll W$ nonzero entries. Calculating the $W \times 1$ projection value vector $\mathbf{v}$ requires two sparse matrix-vector multiplications and takes $\mathcal{O}(MN)$ time.

Finding the maximum requires $\mathbf{W}$ running time; (III) To evaluate one set $\hat{J}_i \leftarrow \{j : \widehat{E}_{i,i} + \widehat{E}_{j,j} - 2\widehat{E}_{i,j} \geq \zeta/2\}$ we need to calculate $\widehat{E}_{i,j}, j = 1, \ldots, W$. This can be viewed as projecting $\widehat{\mathbf{E}}$ along $\mathbf{d} = \mathbf{e}_i$ and takes $\mathcal{O}(MN)$. We also note that the diagonal entries $\mathbf{E}_{w,w}, w = 1, \ldots, W$ can be calculated once using $\mathcal{O}(W)$ time. To sum up, these steps takes $\mathcal{O}(MNP + WP)$ running time.

We then consider the detecting and clustering steps (step 14 to 21 in Alg. 2). We note that all the conditions in Step 17 have been calculated in the previous steps, and recall that the number of novel words are small constant per topic, then, this step will require a running time of $\mathcal{O}(K^2)$.

We last consider the topic estimation steps in Algorithm 3. Here all the corresponding inputs for the linear regression have already been computed in the projection step. Each linear regression has K variables and we upper bound its running time by $\mathcal{O}(K^3)$. Calculating the row-normalization factors $\frac{1}{M}\mathbf{X1}$ requires $\mathcal{O}(MN)$ time. The row and column re-normalization each requires at most $\mathcal{O}(WK)$ running time. Overall, we need a $\mathcal{O}(WK^3 + MN)$ running time.

Other steps are also efficient. Splitting each document into two independent halves takes linear time in N for each document since we can achieve it using random permutation over N items. To generate each random direction $\mathbf{d}_r$ requires $\mathcal{O}(W)$ complexity if we use the spherical Gaussian prior. While we can directly sort the empirical estimated solid angles (in $\mathcal{O}(W\log(W))$ time), we only search for the words with largest solid angles whose number is a constant w.r.t W, therefore it would take only $\mathcal{O}(W)$ time. □

K. Proof of Theorem 3

We focus on the case when the random projection directions are sampled from **any** isotropic distribution. Our proof is not tied to the special form of the distribution; just its isotropic nature. We first provide some useful propositions. We denote by $\mathcal{C}_k$ the set of all novel word of topic k, for $k \in [K]$, and denote by $\mathcal{C}_0$ the set of all non-novel words. We first show,

Proposition 4. *Let $\mathbf{E}_i$ be the i-th row of $\mathbf{E}$. Suppose $\boldsymbol{\beta}$ is separable and $\bar{\mathbf{R}}$ is γ_s-simplicial, then the following is true: For all $k \in [K]$,*

	$\lVert\mathbf{E}_i - \mathbf{E}_j\rVert$	$E_{i,i} - 2E_{i,j} + E_{j,j}$
$i \in \mathcal{C}_k, j \in \mathcal{C}_k$	0	0
$i \in \mathcal{C}_k, j \notin \mathcal{C}_k$	$\geq (1-b)\gamma_s$	$\geq (1-b)^2\gamma_s^2/\lambda_{\max}$

where $b = \max_{j\in\mathcal{C}_0, l} \bar{\beta}_{j,l}$ and $\lambda_{\max} > 0$ is the maximum eigenvalue of $\bar{\mathbf{R}}$

Proof. We focus on the case $k = 1$ since the proofs for other values of k are analogous. Let $\bar{\boldsymbol{\beta}}_i$ be the i-th row vector of matrix $\bar{\boldsymbol{\beta}}$. To show the above results, recall that $\mathbf{E} = \bar{\boldsymbol{\beta}}\bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top$. Then

$$\begin{aligned}&\lVert\mathbf{E}_i - \mathbf{E}_j\rVert = \lVert(\bar{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}}_j)\bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top\rVert\\ &E_{i,i} - 2E_{i,j} + E_{j,j} = (\bar{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}}_j)\mathbf{R}'(\bar{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}}_j)^\top.\end{aligned}$$

It is clear that when $i, j \in \mathcal{C}_1$, i.e., they are both novel word for the same topic, $\bar{\boldsymbol{\beta}}_i = \bar{\boldsymbol{\beta}}_j = \mathbf{e}_1$. Hence, $\lVert\mathbf{E}_i - \mathbf{E}_j\rVert = 0$ and $E_{i,i} - 2E_{i,j} + E_{j,j} = 0$. When $i \in \mathcal{C}_1, j \notin \mathcal{C}_1$, we have $\bar{\boldsymbol{\beta}}_i = [1, 0, \ldots, 0]$, $\bar{\boldsymbol{\beta}}_j = [\bar{\beta}_{j,i}, \bar{\beta}_{j,2}, \ldots, \bar{\beta}_{j,K}]$ with $\bar{\beta}_{j,1} < 1$. Then,

$$\begin{aligned}\bar{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}}_j &= [1 - \bar{\beta}_{j,i}, -\bar{\beta}_{j,2}, \ldots, -\bar{\beta}_{j,K}]\\ &= (1 - \bar{\beta}_{j,i})[1, -c_2, \ldots, -c_K] := (1 - \bar{\beta}_{j,i})\mathbf{e}^\top\end{aligned}$$

and $\sum_{l=2}^{K} c_l = 1$. Therefore, defining $\mathbf{Y} := \bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top$, we get

$$\lVert\mathbf{E}_i - \mathbf{E}_j\rVert_2 = (1 - \bar{\beta}_{j,i})\lVert\mathbf{Y}_1 - \sum_{l=2}^{K} c_l\mathbf{Y}_l\rVert_2$$

Noting that $\mathbf{Y}$ is at least γ_s-simplicial, we have $\lVert\mathbf{E}_i - \mathbf{E}_j\rVert_2 \geq (1-b)\gamma_s$ where $b = \max_{j\in\mathcal{C}_0,k}\bar{\beta}_{j,k} < 1$.

Similarly, note that $\lVert\mathbf{e}^\top\bar{\mathbf{R}}\rVert \geq \gamma$ and let $\bar{\mathbf{R}} = \mathbf{U}\Sigma\mathbf{U}^\top$ be its singular value decomposition. If $\lambda_{\max}$ is the maximum eigenvalue of $\bar{\mathbf{R}}$, then we have

$$\begin{aligned}E_{i,i} - 2E_{i,j} + E_{j,j} &= (1 - \bar{\beta}_{j,1})^2(\mathbf{e}^\top\bar{\mathbf{R}})\mathbf{U}\Sigma^{-1}\mathbf{U}^\top(\mathbf{e}^\top\bar{\mathbf{R}})^\top\\ &\geq (1-b)^2\gamma_s^2/\lambda_{\max}.\end{aligned}$$

The inequality in the last step follows from the observation that $\mathbf{e}^\top\mathbf{R}'$ is within the column space spanned by $\mathbf{U}$. □

The results in Proposition 4 provide two statistics for identifying novel words of the same topic, $\lVert\mathbf{E}_i - \mathbf{E}_j\rVert$ and $E_{i,i} - 2E_{i,j} + E_{j,j}$. While the first is straightforward, the latter is efficient to calculate in practice with better computational complexity. Specifically, its empirical version, the set $\mathcal{J}_i$ in Algorithm 2

$$\mathcal{J}_i = \{j : \widehat{E}_{i,i} - \widehat{E}_{i,j} - \widehat{E}_{j,i} + \widehat{E}_{j,j} \geq d/2\}$$

can be used to discover the set of novel words of the same topics asymptotically. Formally,

Proposition 5. *If $\lVert\widehat{\mathbf{E}} - \mathbf{E}\rVert_\infty \leq (1-b)^2\gamma_s^2/8\lambda_{\max}$, then,*

1) *For a novel word $i \in \mathcal{C}_k$, $\mathcal{J}_i = \mathcal{C}_k^c$*
2) *For a non-novel word $j \in \mathcal{C}_0$, $\mathcal{J}_i \supset \mathcal{C}_k^c$*

Now we start to show that Algorithm 2 can detect all the novel words of the K distinct rankings consistently. As illustrated in Lemma 8, we detect the novel words by ranking ordering the solid angles q_i. We denote the minimum solid angle of the K extreme points by $q_\wedge$. Our proof is to show that the estimated solid angle in Eq (5),

$$\hat{p}_i = \frac{1}{P}\sum_{r=1}^{P}\mathbb{I}\{\forall j \in \mathcal{J}_i,\ \widehat{\mathbf{E}}_j\mathbf{d}^r \leq \widehat{\mathbf{E}}_i\mathbf{d}^r\} \tag{17}$$

converges to the ideal solid angle

$$q_i = \Pr\{\forall j \in \mathcal{S}(i), (\mathbf{E}_i - \mathbf{E}_j)\mathbf{d} \geq 0\} \tag{18}$$

as $M, P \to \infty$. $\mathbf{d}^1, \ldots, \mathbf{d}^P$ are iid directions drawn from a isotropic distribution. For a novel word $i \in \mathcal{C}_k, k = 1, \ldots, K$, let $\mathcal{S}(i) = \mathcal{C}_k^c$, and for a non-novel word $i \in \mathcal{C}_0$, let $\mathcal{S}(i) = \mathcal{C}_0^c$.

To show the convergence of $\hat{p}_i$ to p_i, we consider an intermediate quantity,

$$p_i(\widehat{\mathbf{E}}) = \Pr\{\forall j \in \mathcal{J}_i, (\widehat{\mathbf{E}}_i - \widehat{\mathbf{E}}_j)\mathbf{d} \geq 0\}$$

First, by Hoeffding's lemma, we have the following result.

Proposition 6. $\forall t \geq 0, \forall i,$

$$\Pr\{|\hat{p}_i - p_i(\widehat{\mathbf{E}})| \leq t\} \geq 2\exp(-2Pt^2) \tag{19}$$

Next we show the convergence of $p_i(\widehat{\mathbf{E}})$ to solid angle q_i:

Proposition 7. *Consider the case when* $\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \frac{d}{8}$ *and* $\bar{\mathbf{R}}$ *is* γ_s*-simplicial. If* i *is a novel word, then,*

$$q_i - p_i(\widehat{\mathbf{E}}) \leq \frac{W\sqrt{W}}{\pi d_2}\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty$$

Similarly, if j *is a non-novel word, we have,*

$$p_j(\widehat{\mathbf{E}}) - q_i \leq \frac{W\sqrt{W}}{\pi d_2}\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty$$

where $d_2 \triangleq (1-b)\gamma_s$, $d = (1-b)^2\gamma_s^2/\lambda_{\max}$.

Proof. First note that, by the definition of $\mathcal{J}_i$ and Proposition 4, if $\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \frac{d}{8}$, then, for a novel word $i \in \mathcal{C}_k$, $\mathcal{J}_i = \mathcal{S}(i)$. And for a non-novel word $i \in \mathcal{C}_0$, $\mathcal{J}_i \supseteq \mathcal{S}(i)$. For convenience, let

$$A_j = \{\mathbf{d} : (\widehat{\mathbf{E}}_i - \widehat{\mathbf{E}}_j)\mathbf{d} \geq 0\} \quad A = \bigcap_{j \in \mathcal{J}_i} A_j$$

$$B_j = \{\mathbf{d} : (\mathbf{E}_i - \mathbf{E}_j)\mathbf{d} \geq 0\} \quad B = \bigcap_{j \in \mathcal{S}(i)} B_j$$

For i being a novel word, we consider

$$q_i - p_i(\widehat{\mathbf{E}}) = \Pr\{B\} - \Pr\{A\} \leq \Pr\{B \bigcap A^c\}$$

Note that $\mathcal{J}_i = \mathcal{S}(i)$ when $\|\widehat{\mathbf{E}} - \mathbf{E}\| \leq d/8$,

$$\begin{aligned}
&\Pr\{B \bigcap A^c\} = \Pr\{B \bigcap (\bigcup_{j \in \mathcal{S}(i)} A_j^c)\} \\
&\leq \sum_{j \in \mathcal{S}(i)} \Pr\{(\bigcap_{l \in \mathcal{S}(i)} B_l) \bigcap A_j^c\} \leq \sum_{j \in \mathcal{S}(i)} \Pr\{B_j \bigcap A_j^c\} \\
&= \sum_{j \in \mathcal{S}(i)} \Pr\{(\widehat{\mathbf{E}}_i - \widehat{\mathbf{E}}_j)\mathbf{d} < 0, \text{and } (\mathbf{E}_i - \mathbf{E}_j)\mathbf{d} \geq 0\} \\
&= \sum_{j \in \mathcal{S}(i)} \frac{\phi_j}{2\pi}
\end{aligned}$$

where ϕ_j is the angle between $\mathbf{e}_j = \mathbf{E}_i - \mathbf{E}_j$ and $\widehat{\mathbf{e}}_j = \widehat{\mathbf{E}}_i - \widehat{\mathbf{E}}_j$ for any isotropic distribution on $\mathbf{d}$. Noting that $\phi \leq \tan(\phi)$,

$$\begin{aligned}
\Pr\{B \bigcap A^c\} &\leq \sum_{j \in \mathcal{S}(i)} \frac{\tan(\phi_j)}{2\pi} \leq \sum_{j \in \mathcal{S}(i)} \frac{1}{2\pi} \frac{\|\widehat{\mathbf{e}}_j - \mathbf{e}_j\|_2}{\|\mathbf{e}_j\|_2} \\
&\leq \frac{W\sqrt{W}}{\pi d_2}\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty
\end{aligned}$$

where the last inequality is obtained by the relationship between the ℓ_∞ norm and the ℓ_2 norm, and the fact that for $j \in \mathcal{S}(i)$, $\|\mathbf{e}_j\|_2 = \|\mathbf{E}_i - \mathbf{E}_j\|_2 \geq d_2 \triangleq (1-b)\gamma_s$. Therefore for a novel word i, we have,

$$q_i - p_i(\widehat{\mathbf{E}}) \leq \frac{W\sqrt{W}}{\pi d_2}\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty$$

Similarly for a non-novel word $i \in \mathcal{C}_0$, $\mathcal{J}_i \supseteq \mathcal{S}(i)$,

$$\begin{aligned}
p_i(\widehat{\mathbf{E}}) - q_i &= \Pr\{A\} - \Pr\{B\} = \Pr\{A \bigcap B^c\} \\
&\leq \sum_{j \in \mathcal{S}(i)} \Pr\{(\bigcap_{l \in \widehat{\mathcal{S}}(i)} A_l) \bigcap B_j^c\} \\
&\leq \sum_{j \in \mathcal{S}(i)} \Pr\{A_j \bigcap B_j^c\} \leq \frac{W\sqrt{W}}{\pi d_2}\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty
\end{aligned}$$

□

A direct implication of Proposition 7 is,

Proposition 8. $\forall \epsilon > 0$*, let* $\rho = \min\{\frac{d}{8}, \frac{\pi d_2 \epsilon}{W^{1.5}}\}$*. If* $\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho$*, then,* $q_i - p_i(\widehat{\mathbf{E}}) \leq \epsilon$ *for a novel word* i *and* $p_j(\widehat{\mathbf{E}}) - q_j \leq \epsilon$ *for a non-novel word* j*.*

We now prove Theorem 3. In order to correctly detect all the novel words of K distinct topics, we decompose the error event to be the union of the following two types,

1) *Sorting error*, i.e., $\exists i \in \bigcup_{k=1}^K \mathcal{C}_k, \exists j \in \mathcal{C}_0$ such that $\hat{p}_i < \hat{p}_j$. This event is denoted as $A_{i,j}$ and let $A = \bigcup A_{i,j}$.
2) *Clustering error*, i.e., $\exists k, \exists i, j \in \mathcal{C}_k$ such that $i \notin \mathcal{J}_j$. This event is denoted as $B_{i,j}$ and let $B = \bigcup B_{i,j}$

We point out that the event A, B are different from the notations we used in Proposition 7. According to Proposition 8, we also define $\rho = \min\{\frac{d}{8}, \frac{\pi d_2 q_\wedge}{4W^{1.5}}\}$ and the event that $C = \{\|\mathbf{E} - \widehat{\mathbf{E}}\|_\infty \geq \rho\}$. We note that $B \subsetneq C$.

Therefore,

$$\begin{aligned}
Pe &= \Pr\{A \bigcup B\} \leq \Pr\{A \bigcap C^c\} + \Pr\{C\} \\
&\leq \sum_{i \text{ } novel, j \text{ } non-novel} \Pr\{A_{i,j} \bigcap B^c\} + \Pr\{C\} \\
&\leq \sum_{i,j} \Pr(\hat{p}_i - \hat{p}_j < 0 \bigcap \|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \geq \rho) \\
&\quad + \Pr(\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty > \rho)
\end{aligned}$$

The second term can be bound by Lemma 5. Now we focus on the first term. Note that

$$\begin{aligned}
\hat{p}_i - \hat{p}_j &= \hat{p}_i - \hat{p}_j - p_i(\widehat{\mathbf{E}}) + p_i(\widehat{\mathbf{E}}) \\
&\quad - q_i + q_i - p_j(\widehat{\mathbf{E}}) + p_j(\widehat{\mathbf{E}}) - q_j + q_j \\
&= \{\hat{p}_i - p_i(\widehat{\mathbf{E}})\} + \{p_i(\widehat{\mathbf{E}}) - q_i\} \\
&\quad + \{p_j(\widehat{\mathbf{E}}) - \hat{p}_j\} + \{q_j - p_j(\widehat{\mathbf{E}})\} \\
&\quad + q_i - q_j
\end{aligned}$$

and the fact that $q_i - q_j \geq q_\wedge$, then,,

$$\begin{aligned}
&\Pr(\hat{p}_i < \hat{p}_j \bigcap \|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho) \\
\leq\ &\Pr(p_i(\widehat{\mathbf{E}}) - \hat{p}_i \geq q_\wedge/4) + \Pr(\hat{p}_j - p_j(\widehat{\mathbf{E}}) \geq q_\wedge/4) \\
&+ \Pr(q_i - p_i(\widehat{\mathbf{E}}) \geq q_\wedge/4) \bigcap \|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho) \\
&+ \Pr(p_j(\widehat{\mathbf{E}}) - q_j \geq q_\wedge/4) \bigcap \|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho) \\
\leq\ &2\exp(-Pq_\wedge^2/8) \\
&+ \Pr(q_i - p_i(\widehat{\mathbf{E}}) \geq q_\wedge/4) \bigcap \|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho) \\
&+ \Pr(p_j(\widehat{\mathbf{E}}) - q_j \geq q_\wedge/4) \bigcap \|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho)
\end{aligned}$$

The last equality is by Proposition 6. For the last two terms, by Proposition 8 is 0. Therefore, applying Lemma 5 we obtain,

$$Pe \leq 2W^2\exp(-Pq_\wedge^2/8) + 8W^2\exp(-\rho^2\eta^4 MN/20)$$

And this concludes Theorem 3.

L. Proof of Theorem 4

Without loss of generality, let $1, \dots, K$ be the novel words of topic 1 to K. We first consider the solution of the constrained linear regression. To simplify the notation, we denote $\mathbf{E}_i = [E_{i,1}, \dots, E_{i,K}]$ are the first K entries of a row vector without the super-scripts as in Algorithm 3.

Proposition 9. *Let $\bar{\mathbf{R}}$ be γ_a-affine-independent. The solution to the following optimization problem*

$$\widehat{\mathbf{b}}^* = \arg\min_{b_j \geq 0, \sum b_j = 1} \|\widehat{\mathbf{E}}_i - \sum_{j=1}^{K} b_j \widehat{\mathbf{E}}_j\|$$

converges to the i-th row of $\bar{\boldsymbol{\beta}}$, $\bar{\boldsymbol{\beta}}_i$, as $M \to \infty$. Moreover,

$$\Pr(\|\widehat{\mathbf{b}}^* - \bar{\boldsymbol{\beta}}_i\|_\infty \geq \epsilon) \leq 8W^2 \exp(-\frac{\epsilon^2 MN\gamma_a^2\eta^4}{320K})$$

where η is define the same as in Lemma 5.

Proof. We note that $\bar{\boldsymbol{\beta}}_i$ is the optimal solution to the following problem with ideal word co-occurrence statistics

$$\mathbf{b}^* = \arg\min_{b_j \geq 0, \sum b_j = 1} \|\mathbf{E}_i - \sum_{j=1}^{K} b_j \mathbf{E}_j\|$$

Define $f(\mathbf{E}, \mathbf{b}) = \|\mathbf{E}_i - \sum_{j=1}^{K} b_j \mathbf{E}_j\|$ and note the fact that $f(\mathbf{E}, \mathbf{b}^*) = 0$. Let $\mathbf{Y} = [\mathbf{E}_1^\top, \dots, \mathbf{E}_K^\top]^\top$. Then,

$$\begin{aligned} f(\mathbf{E}, \mathbf{b}) - f(\mathbf{E}, \mathbf{b}^*) &= \|\mathbf{E}_i - \sum_{j=1}^{K} b_j \mathbf{E}_j\| - 0 \\ &= \|\sum_{j=1}^{K} (b_j - b_j^*) \mathbf{E}_j\| = \sqrt{(\mathbf{b} - \mathbf{b}^*)\mathbf{Y}\mathbf{Y}^\top(\mathbf{b} - \mathbf{b}^*)^\top} \\ &\geq \|\mathbf{b} - \mathbf{b}^*\|\gamma_a \end{aligned}$$

The last equality is true by the definition of affine-independence. Next, note that,

$$\begin{aligned} |f(\mathbf{E}, \mathbf{b}) - f(\widehat{\mathbf{E}}, \mathbf{b})| &\leq \|\mathbf{E}_i - \widehat{\mathbf{E}}_i + \sum b_j(\widehat{\mathbf{E}}_j - \mathbf{E}_j)\| \\ &\leq \|\mathbf{E}_i - \widehat{\mathbf{E}}_i\| + \sum b_j \|\widehat{\mathbf{E}}_j - \mathbf{E}_j\| \\ &\leq 2 \max_w \|\widehat{\mathbf{E}}_w - \mathbf{E}_w\| \end{aligned}$$

Combining the above inequalities, we obtain,

$$\begin{aligned} \|\widehat{\mathbf{b}}^* - \mathbf{b}^*\| &\leq \frac{1}{\gamma_a}\{f(\mathbf{E}, \widehat{\mathbf{b}}^*) - f(\mathbf{E}, \mathbf{b}^*)\} \\ &= \frac{1}{\gamma_a}\{f(\mathbf{E}, \widehat{\mathbf{b}}^*) - f(\widehat{\mathbf{E}}, \widehat{\mathbf{b}}^*) + f(\widehat{\mathbf{E}}, \widehat{\mathbf{b}}^*) \\ &\quad - f(\widehat{\mathbf{E}}, \mathbf{b}^*) + f(\widehat{\mathbf{E}}, \mathbf{b}^*) - f(\mathbf{E}, \mathbf{b}^*)\} \\ &\leq \frac{1}{\gamma_a}\{f(\mathbf{E}, \widehat{\mathbf{b}}^*) - f(\widehat{\mathbf{E}}, \widehat{\mathbf{b}}^*) + f(\widehat{\mathbf{E}}, \mathbf{b}^*) - f(\mathbf{E}, \mathbf{b}^*)\} \\ &\leq \frac{4K^{0.5}}{\gamma_a}\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \end{aligned}$$

where the last term converges to 0 almost surely. The convergence rate follows directly from Lemma 5. □

We next consider the row renormalization. Let $\hat{\mathbf{b}}^*(i)$ be the optimal solution in Proposition 9 for the i-th word, and consider

$$\widehat{\mathbf{B}}_i := \hat{\mathbf{b}}^*(i)^\top(\frac{1}{M}\mathbf{X}\mathbf{1}_{M\times 1}) \to \boldsymbol{\beta}_i \operatorname{diag}(\mathbf{a}) \qquad (20)$$

To show the convergence rate of the above equation, it is straightforward to apply the result in Lemma 5

Proposition 10. *For the row-scaled estimation $\hat{\mathbf{B}}_i$ as in Eq.* (20)*, we have,*

$$\Pr(|\hat{\mathbf{B}}_{i,k} - \boldsymbol{\beta}_{i,k}a_k| \geq \epsilon) \leq 8W^2 \exp(-\frac{\epsilon^2 MN\gamma_a^2\eta^4}{1280K})$$

Proof. By Proposition 9, we have,

$$\Pr(|\widehat{\mathbf{b}}^*(i)_k - \bar{\boldsymbol{\beta}}_{i,k}| \geq \epsilon/2) \leq 8W^2 \exp(-\frac{\epsilon^2 MN\gamma_a^2\eta^4}{1280K})$$

Recall that in Lemma 5 by McDiarmid's inequality, we have

$$\Pr(|\frac{1}{M}\mathbf{X}\mathbf{1}_{M\times 1} - \mathbf{B}_i\mathbf{a}| \geq \epsilon/2) \leq \exp(-\epsilon^2 MN/2)$$

Therefore,

$$\begin{aligned} &\Pr(|\hat{\mathbf{B}}_{i,k} - \boldsymbol{\beta}_{i,k}a_k| \geq \epsilon) \\ \leq& 8W^2 \exp(-\frac{\epsilon^2 MN\gamma_a^2\eta^4}{1280K}) + \exp(-\epsilon^2 MN/2) \end{aligned}$$

where the second term is dominated by the first term. □

Finally, we consider the column normalization step to remove the effect of $\operatorname{diag}(\mathbf{a})$:

$$\widehat{\boldsymbol{\beta}}_{i,k} := \widehat{\mathbf{B}}_{i,k} / \sum_{w=1}^{W} \widehat{\mathbf{B}}_{w,k} \qquad (21)$$

And $\sum_{w=1}^{W} \widehat{\mathbf{B}}_{w,k} \to \mathbf{a}_k$ for $k = 1, \dots, K$. A worst case analysis on its convergence is,

$$\begin{aligned} \Pr(|\sum_{w=1}^{W} \widehat{\mathbf{B}}_{w,k} - \mathbf{a}_k| > \epsilon) &\leq W \Pr(|\hat{\mathbf{B}}_{i,k} - \boldsymbol{\beta}_{i,k}a_k| \geq \epsilon/W) \\ &\leq 8W^3 \exp(-\frac{\epsilon^2 MN\gamma_a^2\eta^4}{1280W^2K}) \end{aligned}$$

Combining all the result above, we can show $\forall i = 1, \dots, W, \forall k = 1, \dots, K$,

$$\Pr(|\widehat{\boldsymbol{\beta}}_{i,k} - \boldsymbol{\beta}_{i,k}| > \epsilon) \leq 8W^4 K \exp(-\frac{\epsilon^2 MN\gamma_a^2 a_{\min}^2\eta^4}{2560W^2K})$$

where $a_{\min} > 0$ is the minimum value of entries of $\mathbf{a}$. This concludes the result of Theorem 4.

M. Proof of Lemma 9

Proof. We first show that irreducibility implies separability, or equivalently, if the collection is not separable, then it is not irreducible. Suppose that $\{\nu_1, \dots, \nu_K\}$ is not separable. Then there exists some $k \in [K]$ and a $\delta > 0$ such that,

$$\inf_{A:\ \nu_k(A) > 0} \max_{j:\ j \neq k} \frac{\nu_j(A)}{\nu_k(A)} = \delta > 0.$$

Then $\forall A \in \mathcal{F} : \nu_k(A) > 0$, $\max_{j:\ j\neq k} \frac{\nu_j(A)}{\nu_k(A)} \geq \delta$. This implies that $\forall A \in \mathcal{F} : \nu_k(A) > 0$,

$$\sum_{j:\ j\neq k} \nu_j(A) - \delta\nu_k(A) \geq 0.$$

On the other hand, $\forall A \in \mathcal{F} : \nu_k(A) = 0$, we have

$$\sum_{j:\ j\neq k} \nu_j(A) - \delta\nu_k(A) = \sum_{j:\ j\neq k} \nu_j(A) \geq 0.$$

Thus the linear combination $\sum_{j\neq k} \nu_j - \delta\nu_k$ with one strictly negative coefficient $-\delta$ is nonnegative over all measurable A. This implies that the collection of measures $\{\nu_1, \ldots, \nu_K\}$ is not irreducible.

We next show that separability implies irreducibility. If the collection of measures $\{\nu_1, \ldots, \nu_K\}$ is separable, then by the definition of separability, $\forall k$, $\exists A_n^{(k)} \in \mathcal{F}, n = 1, 2, \ldots,$ such that $\nu_k(A_n^{(k)}) > 0$ and $\forall j \neq k$, $\frac{\nu_j(A_n^{(k)})}{\nu_k(A_n^{(k)})} \to 0$ as $n \to \infty$. Now consider any linear combination of measures $\sum_{i=1}^K c_i\nu_i$ which is nonnegative over all measurable sets, i.e., for all $A \in \mathcal{F}$, $\sum_{i=1}^K c_i\nu_i(A) \geq 0$. Then $\forall k = 1, \ldots, K$ and all $n \geq 1$ we have,

$$\sum_{i=1}^{K} c_i\nu_i(A_n^{(k)}) \geq 0$$

$$\Rightarrow \nu_k(A_n^{(k)}) \left(c_k + \sum_{j\neq k} c_j \frac{\nu_j(A_n^{(k)})}{\nu_k(A_n^{(k)})} \right) \geq 0$$

$$\Rightarrow c_k \geq -\sum_{j\neq k} c_j \frac{\nu_j(A_n^{(k)})}{\nu_k(A_n^{(k)})} \to 0 \text{ as } n \to \infty.$$

Therefore, $c_k \geq 0$ for all k and the collection of measures is irreducible. □

REFERENCES

[1] D. Blei, "Probabilistic topic models," *Commun. of the ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[2] W. Ding, M. Rohban, P. Ishwar, and V. Saligrama, "A new geometric approach to latent topic modeling and discovery," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013.

[3] W. Ding, M. H. Rohban, P. Ishwar, and V. Saligrama, "Topic discovery through data dependent and random projections," in *Proc. of the 30th International Conference on Machine Learning*, Atlanta, GA, USA, Jun. 2013.

[4] W. Ding, P. Ishwar, M. H. Rohban, and V. Saligrama, "Necessary and Sufficient Conditions for Novel Word Detection in Separable Topic Models," in *Advances in on Neural Information Processing Systems (NIPS), Workshop on Topic Models: Computation, Application*, Lake Tahoe, NV, USA, Dec. 2013.

[5] W. Ding, M. H. Rohban, P. Ishwar, and V. Saligrama, "Efficient Distributed Topic Modeling with Provable Guarantees," in *Proc. ot the 17th International Conference on Artificial Intelligence and Statistics*, Reykjavik, Iceland, Apr. 2014.

[6] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.

[7] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, "A practical algorithm for topic modeling with provable guarantees," in *Proc. of the 30th International Conference on Machine Learning*, Atlanta, GA, USA, Jun. 2013.

[8] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[9] S. Arora, R. Ge, and A. Moitra, "Learning topic models – going beyond SVD," in *Proc. of the IEEE 53rd Annual Symposium on Foundations of Computer Science*, New Brunswick, NJ, USA, Oct. 2012.

[10] D. Sontag and D. Roy, "Complexity of inference in latent dirichlet allocation," in *NIPS*, 2011, pp. 1008–1016.

[11] T. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, pp. 5228–5235, 2004.

[12] M. Wainwright and M. Jordan, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1-2, pp. 1–305, 2008.

[13] E. Airoldi, D. Blei, E. Erosheva, and S. Fienberg, *Handbook of Mixed Membership Models and Their Applications*. Chapman and Hall/CRC, 2014.

[14] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade, "A tensor approach to learning mixed membership community models," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2239–2312, 2014.

[15] A. Kumar, V. Sindhwani, and P. Kambadur, "Fast conical hull algorithms for near-separable non-negative matrix factorization," in *the 30th Int. Conf. on Machine Learning*, Atlanta, GA, Jun. 2013.

[16] W. Ding, P. Ishwar, and V. Saligrama, "Most large Topic Models are approximately separable," in *ITA*, 2015.

[17] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999, pp. 50–57.

[18] D. Blei and J. Lafferty, "A correlated topic model of science," *The Ann. of Applied Statistics*, vol. 1, no. 1, pp. 17–35, 2007.

[19] H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proc. of the 26th International Conference on Machine Learning*, Montreal, Canada, Jun. 2009.

[20] M. Hoffman, F. R. Bach, and D. M. Blei, "Online learning for latent dirichlet allocation," in *advances in neural information processing systems*, 2010, pp. 856–864.

[21] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[22] A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley, 2009.

[23] B. Recht, C. Re, J. Tropp, and V. Bittorf, "Factoring nonnegative matrices with linear programs," in *Advances in Neural Information Processing Systems 25*, Lake Tahoe, NV, Dec. 2012, pp. 1223–1231.

[24] S. Vavasis, "On the complexity of nonnegative matrix factorization," *SIAM J. on Optimization*, vol. 20, no. 3, pp. 1364–1377, Oct. 2009.

[25] A. Anandkumar, D. Hsu, A. Javanmard, and S. Kakade, "Learning linear bayesian networks with latent variables," in *the 30th Int. Conf. on Machine Learning*, Atlanta, GA, Jun. 2013.

[26] P. Awasthi and A. Risteski, "On some provably correct cases of variational inference for topic models," *arXiv:1503.06567 [cs.LG]*, 2015.

[27] T. Bansal, C. Bhattacharyya, and R. Kannan, "A provable SVD-based algorithm for learning topics in dominant admixture corpus," in *Advances in Neural Information Processing Systems*, 2014, pp. 1997–2005.

[28] J. Boardman, "Automating spectral unmixing of aviris data using convex geometry concepts," in *Proc. Ann. JPL Airborne Geoscience Workshop*, 1993, p. 1114.

[29] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts?" in *Advances in Neural Information Processing Systems 16*. Cambridge, MA: MIT press, 2004, pp. 1141–1148.

[30] N. Gillis and S. A. Vavasis, "Fast and robust recursive algorithms for separable nonnegative matrix factorization," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 36, no. 4, pp. 698–714, 2014.

[31] V. Farias, S. Jagabathula, and D. Shah, "A data-driven approach to modeling choice," in *Advances in Neural Information Processing Systems*, Vancouver, Canada, Dec. 2009.

[32] W. Ding, P. Ishwar, and V. Saligrama, "A Topic Modeling approach to Ranking," in *Proc. ot the 18th International Conference on Artificial Intelligence and Statistics*, San Diago, CA, May 2015.

[33] A. McCallum, "Mallet: A machine learning for language toolkit," 2002, http://mallet.cs.umass.edu.

[34] D. Lewis, Y. Yang, T. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," *J. Mach. Learn. Res.*, vol. 5, pp. 361–397, Dec. 2004.

[35] L. Yao, D. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 937–946.

[36] D. Newman, A. Asuncion, P. Smyth, and M. Welling, "Distributed algorithms for topic models," *The Journal of Machine Learning Research*, vol. 10, pp. 1801–1828, 2009.

[37] A. Asuncion, P. Smyth, and M. Welling, "Asynchronous distributed learning of topic models," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., 2009, pp. 81–88.

[38] G. Blanchard and C. Scott, "Decontamination of mutually contaminated models," in *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, 2014, pp. 1–9.

[39] C. Scott, “A rate of convergence for mixture proportion estimation, with application to learning from noisy labels,” in *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015, pp. 838–846.

[40] W. Ding, P. Ishwar, and V. Saligrama, “Learning mixed membership mallows model from pairwise comparisons,” in *arXiv: 1504.00757 [cs.LG]*, 2015.